Design and data analysis of kinome microarrays

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Brett Trost

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Catalyzed by protein kinases, phosphorylation is the most important post-translational modification in eukaryotes and is involved in the regulation of almost all cellular processes. Investigating phosphorylation events and how they change in response to different biological conditions is integral to understanding cellular signaling processes in general, as well as to defining the role of phosphorylation in health and disease.

A recently-developed technology for studying phosphorylation events is the kinome microarray, which consists of several hundred "spots" arranged in a grid-like pattern on a glass slide. Each spot contains many peptides of a particular amino acid sequence chemically fixed to the slide, with different spots containing peptides with different sequences. Each peptide is a subsequence of a full protein, containing an amino acid residue that is known or suspected to undergo phosphorylation *in vivo*, as well as several surrounding residues. When a kinome microarray is exposed to cell lysate, the protein kinases in the lysate catalyze the phosphorylation of the peptides on the array. By measuring the degree to which the peptides comprising each spot are phosphorylated, insight can be gained into the upregulation or downregulation of signaling pathways in response to different biological treatments or conditions.

There are two main computational challenges associated with kinome microarrays. The first is array design, which involves selecting the peptides to be included on a given array. The level of difficulty of this task depends largely on the number of phosphorylation sites that have been experimentally identified in the proteome of the organism being studied. For instance, thousands of phosphorylation sites are known for human and mouse, allowing considerable freedom to select peptides that are relevant to the problem being examined. In contrast, few sites are known for, say, honeybee and soybean. For such organisms, it is useful to expand the set of possible peptides by using computational techniques to predict probable phosphorylation sites. In this thesis, existing techniques for the computational prediction of phosphorylation sites are reviewed. In addition, two novel methods are described for predicting phosphorylation events in organisms with few known sites, with each method using a fundamentally different approach. The first technique, called PHOSFER, uses a random forest-based machine-learning strategy, while the second, called DAPPLE, takes advantage of sequence homology between known sites and the proteome of interest. Both methods are shown to allow quicker or more accurate predictions in organisms with few known sites than comparable previous techniques. Therefore, the use of kinome microarrays is no longer limited to the study of organisms having many known phosphorylation sites; rather, this technology can potentially be applied to any organism having a sequenced genome. It is shown that PHOSFER and DAPPLE are suitable for identifying phosphorylation sites in a wide variety of organisms, including cow, honeybee, and soybean.

The second computational challenge is data analysis, which involves the normalization, clustering, statistical analysis, and visualization of data resulting from the arrays. While software designed for the analysis of DNA microarrays has also been used for kinome arrays, differences between the two technologies prompted the development of PIIKA, a software package specifically designed for the analysis of kinome microarray

data. By comparing with methods used for DNA microarrays, it is shown that PIIKA improves the ability to identify biological pathways that are differentially regulated in a treatment condition compared to a control condition. Also described is an updated version, PIIKA 2, which contains improvements and new features in the areas of clustering, statistical analysis, and data visualization. Given the previous absence of dedicated tools for analyzing kinome microarray data, as well as their wealth of features, PIIKA and PIIKA 2 represent an important step in maximizing the scientific value of this technology.

In addition to the above techniques, this thesis presents three studies involving biological applications of kinome microarray analysis. The first study demonstrates the existence of "kinotypes"—species- or individual-specific kinome profiles—which has implications for personalized medicine and for the use of model organisms in the study of human disease. The second study uses kinome analysis to characterize how the calf immune system responds to infection by the bacterium *Mycobacterium avium* subsp. *paratuberculosis*. Finally, the third study uses kinome arrays to study parasitism of honeybees by the mite *Varroa destructor*, which is thought to be a major cause of colony collapse disorder.

In order to make the methods described above readily available, a website called the SAskatchewan PHosphorylation Internet REsource (SAPHIRE) has been developed. Located at the URL `http://saphire.usask.ca`, SAPHIRE allows researchers to easily make use of PHOSFER, DAPPLE, and PIIKA 2. These resources facilitate both the design and data analysis of kinome microarrays, making them an even more effective technique for studying cellular signaling.

Acknowledgements

First and foremost, I would like to thank Dr. Tony Kusalik, who has provided me with an incredible amount of guidance, advice, and knowledge over the years—first as an instructor, and then as a supervisor for summer research projects, my Master's, and finally my Ph.D. Few people have as much depth and breadth of knowledge in the field of bioinformatics as Dr. Kusalik, and I feel deeply privileged to have been able to benefit from his expertise. Dr. Kusalik's cheerful demeanour and kind personality also contributed greatly to making my time at the University of Saskatchewan enjoyable (his frequent fill-ups of the lab candy dish and offerings of produce from his garden didn't hurt either). Special thanks also goes to Dr. Scott Napper, whose enthusiasm for research is incredibly contagious, and who has a gift for producing a seemingly never-ending stream of great research ideas. Having such a close wet-lab collaborator was very valuable to me, both in terms of identifying biological questions of interest (that computer science can be used to help answer), and in terms of brainstorming ideas for maximizing the benefit of the fusion between computer science and biology. The other members of my advisory committee—Dr. Ian McQuillan, Dr. Michael Horsch, and Dr. Mik Bickis—also have my gratitude for their invaluable advice, contributions, and support. I would also like to thank Erin Scruten for answering my questions regarding wet-lab aspects of kinome microarray experiments and for helping test PIIKA 2. In addition, I would like to acknowledge my colleagues in the lab, who not only offered plenty of help and advice with respect to research and technical problems, but also friendship, and maybe even a few opportunities for procrastination.

On a more personal level, I would like to acknowledge my parents, Randy and Ruth, and my sister, Kelli, for being the most wonderful family a person could ask for. And last but certainly not least, I want to thank my wife, Chantel, for being the love of my life, and for all the little (and not-so-little) things that she does for me.

Funding for my degree was generously provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) Vanier Canada Graduate Scholarship program, the College of Graduate Studies and Research, the Department of Computer Science, the Government of Saskatchewan, and Dr. Kusalik's NSERC grant.

Contents

List of Tables

List of Figures

List of Abbreviations

ADP	Adenosine diphosphate
ANN	Artificial neural network
A_{ROC}	Area under the receiver operating characteristic curve
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
CCD	Colony collapse disorder
CDK	Cyclin-dependent kinase
cDNA	Complementary deoxyribonucleic acid
CFC	Consensus fuzzy clustering
CFU	Colony forming units
CGI	Common gateway interface
CMGC	A family of protein kinases that includes cyclin-dependent kinases, mitogen-activated protein kinases, glycogen synthase kinases, and cyclin-dependent kinase-like kinases
DNA	Deoxyribonucleic acid
DWV	Deformed wing virus
EBI	European Bioinformatics Institute
ELISA	Enzyme-linked immunosorbent assay
FC	Fold-change
FN	False negative
FP	False positive
FSA	Finite-state automaton
GSK	Glycogen synthase kinase
HCV	Hepatitis C virus
IAPV	Israeli acute paralysis virus
IFN	Interferon
IL	Interleukin
JAK-STAT	Janus kinase-signal transducer and activator of transcription
JD	Johne's disease
KBV	Kashmir bee virus
KEGG	Kyoto Encyclopedia of Genes and Genomes
LPS	Lipopolysaccharide
MAP	*Mycobacterium avium* subsp. *paratuberculosis*
MAPK	Mitogen-activated protein kinase
MAPKK	Mitogen-activated protein kinase kinase
MAPKKK	Mitogen-activated protein kinase kinase kinase
MCC	Matthews correlation coefficient
MLN	Mesenteric lymph node
MPXV	Monkeypox virus
mRNA	Messenger ribonucleic acid
mt	Mitochondrion
NAR	*Nucleic Acids Research*
NCBI	National Center for Biotechnology Information
NPV	Negative predictive value
NSERC	Natural Sciences and Engineering Research Council of Canada
ORA	Over-representation analysis
PBMC	Peripheral blood mononuclear cell
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction

PHB	Per hundred bees
PHOSFER	PHOsphorylation Site FindER
PIIKA	Platform for Intelligent, Integrated Kinome Analysis
PKA	Protein kinase A
PMN	Polymononuclear cell
PPV	Positive predictive value
PSSM	Position-specific scoring matrix
PTM	Post-translational modification
qRT-PCR	Quantitative reverse-transcription polymerase chain reaction
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
SAPHIRE	SAskatchewan PHosphorylation Internet REsource
SD	Standard deviation
SE	Standard error
SGD	*Saccharomyces* Genome Database
SOM	Self-organizing map
SVM	Support vector machine
TAIR	The *Arabidopsis* Information Resource
TLR	Toll-like receptor
TN	True negative
TP	True positive
t-SNE	t-distributed stochastic neighbor embedding
UPGMA	Unweighted pair group method with arithmetic mean
VSH	Varroa sensitive hygiene
WU-BLAST	Washington University basic local alignment search tool

Chapter 1

Introduction

Despite their small size, biological cells are incredibly complex machines, performing numerous functions that enable an organism to respond to changes in its environment. For example, organisms commonly described as "warm-blooded" can respond to a decrease in environmental temperature by increasing their metabolisms in order to generate additional heat. The activation of the immune system when a microbial infection occurs also constitutes a response to an environmental stimulus. Cellular signaling—that is, communication between different parts of the same cell or between different cells—is an integral part of these responses, as it bridges the gap between the detection of an environmental change and the changes in cellular physiology needed to respond to it. Defects in signaling are responsible for many serious human diseases, such as cancer and autoimmunity. Therefore, studying cellular signaling has the potential to enhance our understanding of cellular physiology in general and to provide insight into the pathogenesis of, and potential treatments for, various diseases.

The most common mechanism of cellular signaling is protein phosphorylation, in which a phosphate group is attached to a protein in order to modify its behaviour. For instance, phosphorylating a protein may cause it to be activated or inactivated, or it may affect which other molecules interact with that protein. The chemical reaction needed to phosphorylate a protein is catalyzed by enzymes called protein kinases. Phosphorylation is involved in the regulation of almost all cellular processes, so understanding how various stimuli affect the phosphorylation of proteins can contribute significantly to our overall understanding of cellular physiology.

In the past, phosphorylation was studied primarily using low-throughput biological techniques—that is, techniques that provide information on only one protein at a time. However, cellular signaling networks are very complex, with one stimulus potentially giving rise to an entire chain of phosphorylation reactions and other biological events. Thus, understanding these networks as a whole requires the use of techniques that provide information on many proteins at once.

The kinome microarray is a technology for investigating phosphorylation-mediated cellular signaling in a high-throughput manner. Kinome arrays are glass slides containing short peptides, usually 15 amino acids in length. Each peptide has two essential properties: first, it is a subsequence of a full protein produced by the organism of interest, and second, its central residue is a phosphorylation site (that is, a residue that is known or suspected to be phosphorylated *in vivo*). Each "spot" on the array contains many copies of a peptide of a particular sequence. In experiments involving kinome microarrays, cells are extracted from the

organism of interest and broken open. The array is then exposed to the contents of the cells for a period of time, during which the protein kinases from the cells catalyze the phosphorylation of the peptides on the array. The degree to which the peptides of a particular sequence are phosphorylated can then be measured. Since a kinome microarray can contain hundreds of unique peptides, the resulting information can provide substantial insight into the global state of the cells' signaling networks.

One of the challenges encountered in studies involving kinome microarrays is array design, which involves choosing appropriate peptides to include on the arrays. As mentioned above, the peptides on an array must be derived from proteins produced by the organism being studied, and must contain an amino acid residue that is known or suspected to be phosphorylated. Unfortunately, while many phosphorylation sites have been experimentally identified in some organisms, few sites have been identified for many other organisms. If the organism of interest is in the latter group, it is difficult to design kinome arrays suitable for studying it. Table 1.1 shows how many known phosphorylation sites from various organisms are found in the phosphorylation site databases PhosphoSitePlus [Hornbeck et al., 2004, 2012], Phospho.ELM [Diella et al., 2004, 2008, Dinkel et al., 2011], P^3DB [Gao et al., 2009b, Yao et al., 2012], and PhosphoGRID [Stark et al., 2010, Sadowski et al., 2013]. This table suggests that designing arrays for organisms like human, mouse, and rat is relatively easy, as there are many experimentally-determined phosphorylation sites from which to choose. Conversely, designing arrays for some other organisms is more difficult because of a paucity of known phosphorylation sites. For instance, fewer than 200 sites are known for each of pig, corn, frog, hamster, dog, potato and sheep. This is fewer than the number of unique peptides that are typically included on an array.

In the absence of a sufficient number of experimentally-determined phosphorylation sites in the organism of interest, phosphorylation sites must be predicted computationally. Many machine-learning techniques for predicting phosphorylation sites have been described (see Chapter 3 for a comprehensive review); however, most of these predict only for mammalian phosphorylation sites, and none concentrate on the prediction of sites in organisms for which few sites have been experimentally determined—exactly the types of organisms for which prediction is most useful. Therefore, one of the goals of this thesis was to develop computational tools for predicting phosphorylation sites in organisms with few known sites. If phosphorylation sites in these organisms can be accurately predicted, then kinome microarrays can be designed for studying them. This thesis presents two methods for the computational prediction of phosphorylation sites: PHOsphorylation Site FindER (PHOSFER), which uses a machine-learning approach, and DAPPLE, which uses a homology-based approach.

Another challenge encountered when using kinome arrays is data analysis. Given the amount of information they produce, appropriate techniques for clustering, statistical analysis, and data visualization are needed in order to make sense of the data and to identify biologically meaningful patterns. Although studies involving the use of kinome arrays have used software designed for DNA microarrays (see Chapter 7 for a discussion of these), differences between the two technologies prompted us to develop software specifically designed for the analysis of kinome arrays. This thesis presents two iterations of this software. The first,

Table 1.1: Number of phosphorylation sites for each organism in each major phosphorylation site database. An organism is listed only if there is a database containing at least 10 sites from it. The scientific name is given for each organism, along with its common name in parentheses (when applicable). When more than one database contains phosphorylation sites from a given organism, the sum of those numbers is not necessarily meaningful, as some overlap exists among the databases.

Organism	PhosphoSitePlus	Phospho.ELM	P^3DB	PhosphoGRID
Homo sapiens (human)	160,735	37,145	0	0
Mus musculus (mouse)	79,435	8,038	0	0
Rattus norvegicus (rat)	15,368	562	0	0
Medicago truncatula	0	0	15,683	0
Arabidopsis thaliana	0	0	15,465	0
Oryza sativa (rice)	0	0	12,317	0
Saccharomyces cerevisiae (yeast)	0	57	0	6,440
Caenorhabditis elegans	0	5,651	0	0
Drosophila melanogaster (fruit fly)	10	5,342	0	0
Glycine max (soybean)	0	1	2,739	0
Vitis vinifera (grape)	0	0	862	0
Brassica napus (rapeseed)	0	0	818	0
Bos taurus (cow)	505	188	0	0
Gallus gallus (chicken)	364	120	0	0
Oryctolagus cuniculus (rabbit)	180	92	0	0
Sus scrofa (pig)	138	18	0	0
Zea mays (corn)	0	3	115	0
Xenopus laevi (frog)	34	40	0	0
Mesocricetus auratus (hamster)	41	23	0	0
Canis lupus familiaris (dog)	43	5	0	0
Solanum tuberosum (potato)	0	0	33	0
Ovis aries (sheep)	12	12	0	0
Torpedo californica (pacific electric ray)	2	12	0	0
Clupea pallasii (pacific herring)	0	10	0	0
Capra aegagrus hircus (goat)	10	0	0	0
Nicotiana tabacum (tobacco)	0	0	10	0

called Platform for Intelligent, Integrated Kinome Analysis (PIIKA), has facilities for cluster analysis, statistical comparisons between samples, and data visualization. The second, PIIKA 2, contains many additions and improvements over the original PIIKA, including a web-based interface.

In order to make the above-described tools as widely accessible as possible, a website was created in order to host them. This website, called the SAskatchewan PHosphorylation Internet REsource (SAPHIRE), can be found at `http://saphire.usask.ca`.

Of course, efforts to facilitate the design and data analysis of kinome microarrays are of little use if the arrays are not actually used to study real biological systems. Thus, this thesis presents three such studies. The first establishes the existence of "kinotypes", which are species- or individual-specific profiles of protein kinase activity. The fact that a given species has different basal kinase activities than another species has implications for the use of model organisms for investigating human disease, as these differences must be taken into account when drawing conclusions about the applicability of the findings to human. Additionally, the fact that individuals exhibit different phosphorylation patterns has implications for personalized medicine; it is possible, for instance, that a given treatment might be effective only for individuals exhibiting particular patterns of protein kinase activity. The second study examines the kinomic responses in calves to infection by the bacterium *Mycobacterium avium* subsp. *paratuberculosis* (MAP), which causes an ailment called Johne's Disease (JD). This study shows that the phosphorylation patterns of calves infected with MAP related to whether a cell-mediated immune response or an antibody response was elicited—with the former being more effective at clearing the infection than the latter. The results of this study could lead to treatments for JD in cattle, with the general strategy of steering the immune response from the antibody type to the cell-mediated type. The third study concerns parasitism of honeybees (*Apis mellifera*) by the mite *Varroa destructor*, which has detrimental effects on honeybee colonies and is thought to be partially responsible for the significant declines in honeybee populations observed over the past few years. By analyzing kinome responses in infected or uninfected bees that are either resistant or susceptible to Varroa, it is shown that Varroa infestation may cause changes in innate immune function in the susceptible bees that make them more vulnerable to viral infections.

This thesis is organized as follows. Chapter 2 contains the background information needed to understand this document. Chapters 3–11 are the main body of the thesis, with each chapter representing a self-contained body of work that contributes to the goals described above. Chapter 12 contains some additional discussion beyond that contained in the individual chapters, as well as some concluding remarks, while Chapter 13 presents ideas for future work.

Each main chapter (Chapters 3–11) includes a short discussion of how the work fits in with the thesis as a whole. Except for Chapter 6, each of these chapters contains a paper that has been published in a peer-reviewed journal. Appendix A contains information about the permissions required to reproduce these papers here. The remaining appendices contain supplementary material published along with the published papers. All references in this document, including those cited in appendices, can be found following the end

of Chapter 13.

The papers composing this thesis are divided into three groups. The first group contains papers related to the design of kinome microarrays. Specifically, Chapter 3 contains a comprehensive review of the literature relating to the computational prediction of phosphorylation sites, while Chapters 4 and 5 describe the PHOSFER and DAPPLE methods, respectively. Chapter 6 presents a case study in the use of DAPPLE to design a honeybee-specific kinome array. The second group of chapters contains papers relating to the analysis of kinome microarray data. Chapter 7 describes the original version of PIIKA, while its successor, PIIKA 2, is described in Chapter 8. The third group contains papers describing the application of the arrays to biological problems. Chapter 9 demonstrates the existence of species- and individual-specific kinotypes, while Chapter 10 describes the use of kinome microarrays to study MAP infection in calves. Finally, Chapter 11 describes the application of a honeybee-specific kinome array to study parasitism by the Varroa mite.

Overall, this thesis makes contributions to three areas of kinome microarray analysis. First, it presents two new methods (PHOSFER and DAPPLE) for predicting phosphorylation sites in species with few known sites, facilitating the design of kinome arrays that are specific to such species. Using soybean as a test case, it is shown that PHOSFER has greater predictive accuracy than previous machine-learning methods for predicting plant phosphorylation sites. In addition, it is shown that DAPPLE can successfully predict phosphorylation sites in a diverse range of organisms, including cow and honeybee. PHOSFER and DAPPLE expand the number of species to which kinome microarray analysis may be applied, significantly increasing the value of this technology, and are also applicable to any area of research that would benefit from the prediction of phosphorylation sites. Second, this thesis presents PIIKA and PIIKA 2, the first software pipelines specifically designed for the analysis of kinome microarray data. They provide a complete suite of data analysis functions, including normalization, quantification of reproducibility, statistical comparisons among samples, clustering, and visualization. In particular, it is shown that PIIKA facilitates more accurate identification of differentially regulated signaling pathways than software designed for DNA microarrays. Finally, this thesis presents three studies that apply the previously-mentioned methods to actual biological problems, showing that they are indeed of practical value. Collectively, the methods and studies presented here allow kinome microarrays to be applied to a greater number of species, and improve the ability to analyze data resulting from the arrays. As such, they should make kinome microarrays an even more valuable tool for addressing important biological questions.

Chapter 2

Background

This chapter contains background information necessary to understand the rest of this document, and is divided into three major sections. Section 2.1 contains an introduction to relevant biology concepts. Section 2.2 discusses the kinome microarray, which is the technology that is the focus of this thesis. Finally, Section 2.3 describes relevant computer science concepts.

2.1 Biology concepts

This thesis concentrates on the design and data analysis of kinome microarrays. The purpose of this section is to give enough biology background to understand what kinome microarrays are, how they relate to similar technologies, how experiments using them are performed, and what biological inferences can be drawn from the resultant data. Specifically, Section 2.1.1 introduces nucleic acids, while Section 2.1.2 discusses proteins and their biological role. Section 2.1.3 explains the concept of homology, which is concerned with evolutionary relationships among proteins. Section 2.1.4 covers phosphorylation, which is the addition of a certain chemical group to a protein after it has been synthesized. The enzymes that catalyze phosphorylation reactions, called protein kinases, are discussed in Section 2.1.5. The phosphorylation of proteins by protein kinases is an important part of cellular signaling pathways, which are described in Section 2.1.6.

2.1.1 Nucleic acids

Nucleic acids are molecules used for the storage or transmission of instructions required for the functioning of living organisms. One type of nucleic acid, deoxyribonucleic acid (DNA), is responsible for storing these instructions. DNA consists of sequences of four bases called adenine, cytosine, guanine, and thymine, which are abbreviated by the letters A, C, G, and T, respectively. The genomes of even the simplest organisms contain hundreds of thousands of these bases, while the human genome consists of approximately 3.2 billion [Lander et al., 2001].

DNA is double-stranded, with the bases on one strand forming complementary pairs (facilitated by interactions among the atoms comprising the bases) with the bases on the other strand; adenine is complementary to thymine, while cytosine is complementary to guanine. A single DNA molecule has directionality; one end is called the $5'$ end, while the other is called the $3'$ end. The strands in double-stranded DNA run in opposite

$5' - \texttt{ATGCAATCGGCGCTAGAGCTCGC} - 3'$
$3' - \texttt{TACGTTAGCCGCGATCTCGAGCG} - 5'$

Figure 2.1: A double-stranded sequence of DNA.

directions, so the base at the $5'$ end of one strand is paired with the base at the $3'$ end of the other strand. Given that DNA strands run in opposite directions, and that pairs of bases are complementary, one strand of DNA is said to be the reverse complement of the other strand. An example of a double-stranded sequence of DNA is shown in Figure 2.1.

A gene is a functional unit of DNA, and serves as a template for the synthesis of another type of nucleic acid, ribonucleic acid (RNA). RNA is similar to DNA, with a few small but significant chemical differences. One difference is that in RNA a base called uracil (abbreviated U) is used instead of thymine. Like thymine, uracil pairs with adenine. The process of producing an RNA molecule from a gene is called transcription. The RNA produced by using a segment of DNA as a template is identical in sequence to one of the strands (the coding strand) except that thymine is replaced with uracil, and is the reverse complement of the other strand (the template strand). For instance, if the top strand in Figure 2.1 is the coding strand and the bottom strand is the template strand, then the transcribed RNA would have the sequence $5' - \texttt{AUGCAAUCGGCGCUAGAGCUCGC} - 3'$.

While some RNA molecules have structural or regulatory roles within the cell, most genes code for messenger RNAs (mRNAs), which serve as templates for the synthesis of proteins (described further in Section 2.1.2). The process of producing a protein using an mRNA molecule as a template is called translation. The overall process of producing proteins from the information contained in genes is called gene expression. Some genes are always expressed at approximately the same level, while the expression of other genes may increase or decrease depending on various factors. For instance, genes encoding proteins involved in carbohydrate metabolism may be expressed at a higher level after a meal.

2.1.2 Proteins

Proteins are large molecules that perform a substantial portion of the operations required for the functioning of a cell. They are composed of smaller molecules called amino acids joined together in a specific sequence. An amino acid that is part of a protein is commonly called an amino acid residue or simply a residue, since the joining process entails removing part of the amino acid. Twenty different amino acids are commonly found in proteins, with a portion of their structure being common to all amino acids (the backbone) and the other portion being unique to a particular amino acid (the side chain). The chemical properties of a given amino acid depend on its side chain; for instance, some side chains are positively charged, while others are hydrophobic (water hating). Figure 2.2A illustrates the general structure of an amino acid, while parts B, C and D contain structures of specific amino acids. The process of two amino acids being joined together

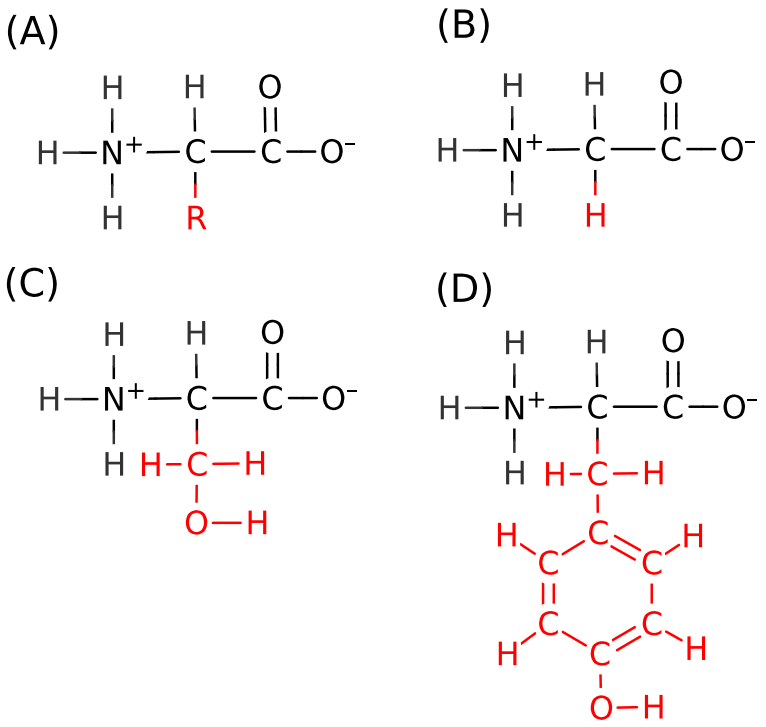


Figure 2.2: Structures of amino acids. Letters represent atoms (N for nitrogen, C for carbon, O for oxygen, and H for hydrogen), while lines represent chemical bonds. Portions common to all amino acids are shown in black, while portions specific to particular amino acids are shown in red. (A) General structure of an amino acid, where R represents the portion of the molecule that differs among the 20 amino acids found in proteins. (B) The structure of glycine, the simplest possible amino acid. (C) The structure of serine. (D) The structure of tyrosine.

is illustrated in Figure 2.3. Proteins can range in length from a few amino acids to hundreds of amino acids, although very short sequences (say, less than 40 or 50 amino acids) are usually referred to as peptides rather than proteins. Amino acids are commonly denoted by three- or one-letter codes; Table 2.1 contains a list of the amino acids along with their abbreviations. These one-letter codes should not be confused with those used for nucleic acids—for instance, G stands for glycine in the context of proteins and guanine in the context of nucleic acids. The sequence of a 5-mer (a peptide containing 5 amino acid residues) containing, in order, leucine, methionine, proline, tyrosine, and alanine would be written as LMPYA (or less commonly, Leu-Met-Pro-Tyr-Ala). Proteins are directional; thus, LMPYA is not the same molecule as AYPML. The beginning of a protein is called its N-terminus, while the end of a protein is called its C-terminus. Individual residues in a protein are often referred to by their position in the sequence and the residue found at that position; for instance, Y77 means that the 77^{th} residue in the sequence is tyrosine.

The amino acid sequence of a particular protein is determined by the DNA sequence of its corresponding gene. A protein's sequence is the primary determinant of its three-dimensional shape, which in turn dictates its biological function. Many different proteins are produced by higher organisms; for example, the human

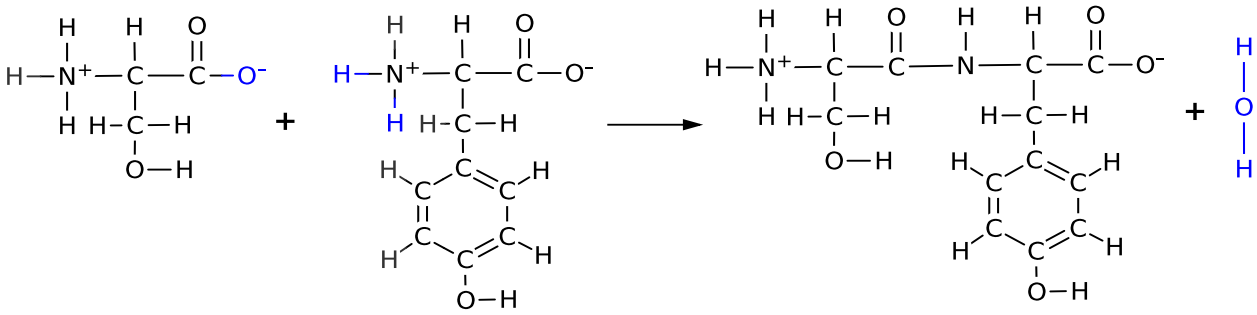


Figure 2.3: The joining of the amino acids serine and tyrosine. The blue atoms are removed as part of the reaction to form H_2O (water).

Table 2.1: List of amino acids and their three- and one-letter codes.

Amino acid	Three-letter code	One-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

genome is estimated to contain more than 20,000 protein-coding genes [ENCODE Project Consortium et al., 2012]. The entire complement of proteins produced by a given organism is called its proteome.

Proteins perform a large variety of biochemical functions. For instance, some proteins assist other proteins in assuming the correct three-dimensional shape, while others act as pores that allow smaller molecules to enter or exit the cell. Many of the structural components of the ribosome, the cellular machine that synthesizes new proteins, are themselves proteins. Finally, most of the chemical reactions that occur inside a cell are catalyzed by proteins called enzymes. A particular class of enzyme called a protein kinase is of particular interest in this thesis, and is described in Section 2.1.5.

2.1.3 Homology

With respect to proteins, homology refers to a similarity in sequence between two proteins arising from common ancestry. Two types of homology are commonly used to describe relationships among proteins: orthology and paralogy.

Orthology refers to the situation in which two proteins with similar or identical functions in two different species descended from a single protein in a common ancestor. For instance, suppose that species A, which encodes a protein called P_A, gives rise to species B and C. Further suppose that species B and C encode proteins called P_B and P_C, respectively, each of which descended from P_A. Then P_B and P_C are called orthologues. Due to evolutionary mutations in their corresponding gene sequences, P_B and P_C may diverge in sequence both from P_A and from each other. More rarely, functional divergence can also occur.

Genes can occasionally become duplicated due to errors in DNA replication. Paralogy refers to the case where two proteins in the same species arose from a single protein due to a gene duplication event. For instance, suppose that the gene encoding protein P_1 is duplicated in some organism, and that the second gene encodes a protein called P_2. Then P_1 and P_2 are called paralogues. Functional divergence is more common in paralogues than in orthologues, with the two proteins usually performing distinct but related functions.

2.1.4 Phosphorylation

After a protein is synthesized, it may undergo one or more post-translational modifications (PTMs). While there are several broad categories of modifications, some of the most common involve the addition of small chemical groups to specific amino acids. One of the most important PTMs—and the one that is of interest in this thesis—is phosphorylation, which involves the transfer of a phosphate group from the molecule adenosine triphosphate (ATP) to a protein. The phosphate group is almost always added to a Ser, Thr, or Tyr residue, although other residues, such as His, can occasionally be phosphorylated. Figure 2.4 illustrates the transfer of a phosphate group from ATP to a Ser residue. Phosphorylating a protein changes its three-dimensional shape, which can have effects like preventing it from being degraded, directing it to a specific location within the cell, or activating, deactivating, or modifying its activity. Phosphorylation is extremely widespread,

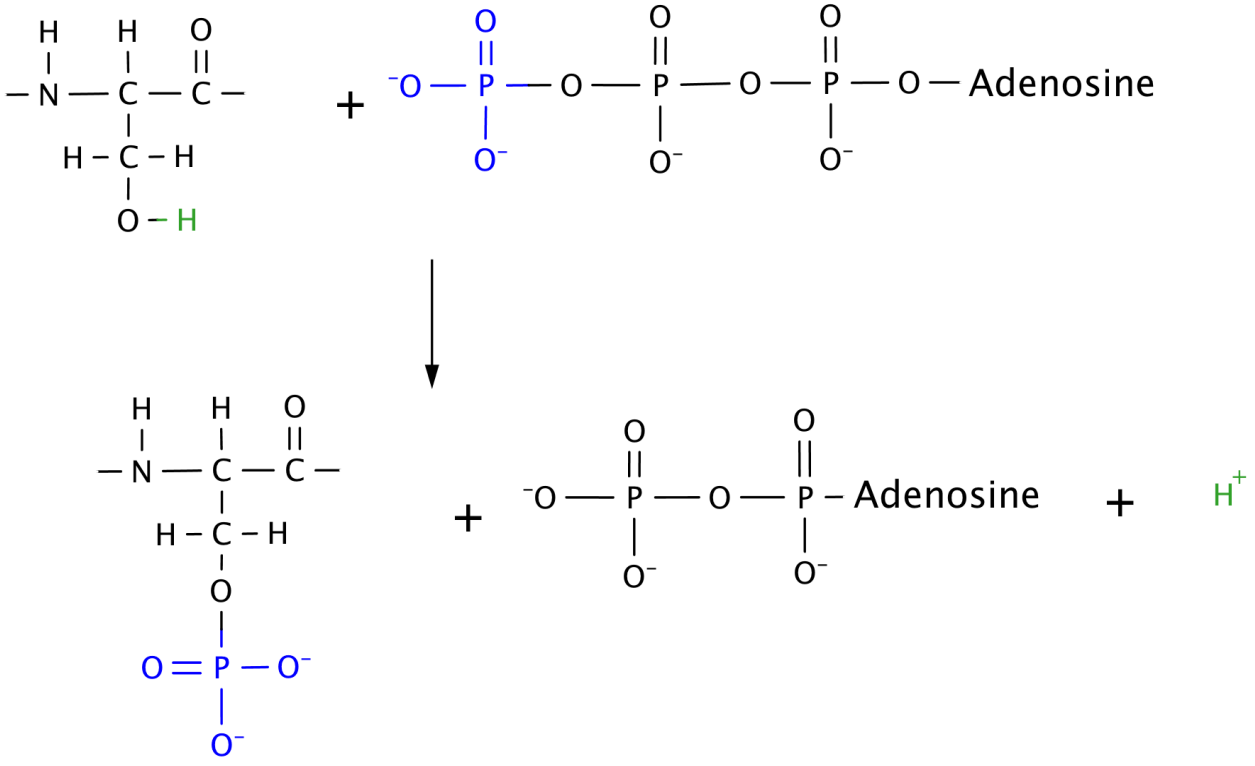


Figure 2.4: The chemical reaction between a serine residue and ATP to form phosphoserine and adenosine diphosphate (ADP). Atoms from ATP that are transferred to the serine residue are shown in blue, while the atom from serine that is transferred to ATP is shown in green.

with one-third of all proteins in the eukaryotic cell estimated to undergo this PTM [Johnson and Hunter, 2005]. The entire complement of phosphorylated proteins in a given organism is called its phosphoproteome. Some proteins have many phosphorylation sites—that is, serine, threonine, or tyrosine residues that can be phosphorylated. For instance, PhosphoSitePlus [Hornbeck et al., 2004, 2012], a database of experimentally verified phosphorylation sites, shows that 12 residues are known to undergo phosphorylation in the protein cyclin-dependent kinase 2 (CDK2).

There are several laboratory techniques for identifying phosphorylation sites in proteins. Some of these allow the analysis of only one protein (or at most a few proteins) at a time, and are thus described as "low-throughput". For instance, a technique called site-directed mutagenesis can be used to provide evidence that the phosphorylation of a particular residue is required for a given protein's activity. This technique involves modifying the DNA sequence of the gene coding for the protein of interest to replace a phosphorylatable amino acid (Ser, Thr, or Tyr) in the protein product with one that cannot be phosphorylated (usually Ala) [e.g., Gu et al., 1992]. If this results in the activation or inactivation of the protein, then that residue is likely a phosphorylation site. Another technique for determining the locations of phosphorylation sites is tryptic phosphopeptide mapping [e.g., Lees et al., 1991]. In this technique, the protein of interest is digested

(broken into smaller peptides) using an enzyme called trypsin, which cleaves proteins at Lys or Arg residues that are not followed by a Pro residue. Because the specificity of trypsin is well-defined, the sequences of the resulting peptides can easily be predicted. The resultant peptides can then be separated from each other, and the subset of peptides that are phosphorylated can be detected. Yet another technique is Edman phosphate-release sequencing [MacDonald et al., 2002], wherein the protein of interest is digested into shorter peptides, and the resulting peptides are subjected to Edman degradation, which involves the cleavage of one amino acid residue at a time from the N-terminus of the peptides. After each cycle, the release of a phosphorylated amino acid can be detected, which allows the positions of the phosphorylated residues in the peptides to be inferred.

Protein kinases and/or their substrates can also be studied using high-throughput techniques, which facilitate the analysis of many molecules simultaneously. One such technique is mass spectrometry, which allows the phosphorylation status of particular amino acid residues to be inferred by measuring the masses of peptides derived from cellular proteins. For example, Ficarro et al. [2002] used mass spectrometry to identify several hundred phosphorylation sites in the yeast proteome, while Nakagami et al. [2010] used this technology to identify thousands of phosphorylated residues in rice proteins. In addition to simply identifying residues that become phosphorylated, it is often of interest to compare the degree to which different proteins are phosphorylated among two or more samples. Although relatively uncommon, mass spectrometry has been used for this type of analysis [Jalal et al., 2007]. For instance, Zheng et al. [2005] used this technique to compare the degree of phosphorylation of many tyrosine phosphorylation sites in cells that were either treated or untreated with a particular immune-related protein. As another example, mass spectrometry was used by Yang et al. [2006] to measure changes in phosphorylation levels after human skin cells were given either low or high doses of radiation. One disadvantage of using mass spectrometry to measure changes in phosphorylation levels is that some proteins are phosphorylated at very low levels in all conditions, making it difficult to distinguish real (albeit small) changes in phosphorylation levels from changes that are merely the result of noise [Jalal et al., 2007].

Another technology that allows the high-throughput analysis of phosphorylation is the kinome microarray. Whereas mass spectrometry measures the phosphoproteome directly, kinome arrays measure the activities of the enzymes that catalyze phosphorylation reactions [Jalal et al., 2007]. Unlike mass spectrometry, kinome microarrays are not well-suited to identifying novel phosphorylation sites, as the technology requires that a set of phosphorylation sites be known in advance. However, kinome arrays are well-suited for measuring changes in phosphorylation patterns among samples, as they do not suffer from the limitation of mass spectrometry mentioned above. As kinome arrays are the focus of this thesis, they are discussed in detail in Section 2.2.

2.1.5 Protein kinases

The phosphorylation of proteins is catalyzed by other proteins called protein kinases. The importance of phosphorylation as a PTM is reflected in the number of protein kinases produced by higher organisms; for

instance, the human genome encodes more than 500 different protein kinases [Manning et al., 2002]. The entire complement of protein kinases encoded by a given organism is called its kinome.

Each kinase has a characteristic recognition pattern, and will catalyze the phosphorylation of a given residue only if the surrounding residues match that pattern [Diks et al., 2007]. For example, one particular class of kinases usually requires that the phosphorylated residue be followed by three Met residues [Ubersax and Ferrell, 2007]. However, recognition patterns are rarely as clear-cut as the aforementioned pattern suggests; in reality, there may be instances where this class of kinases phosphorylates residues not followed by three Met residues, as well as instances of phosphorylatable residues (i.e., Ser, Thr, or Tyr) followed by three Met residues that are not phosphorylated by these kinases. Some kinases catalyze the phosphorylation of only a few different proteins, while others appear to have hundreds of targets [Ubersax and Ferrell, 2007]. At the broadest level, protein kinases can be categorized based on the amino acid residues that they act upon, with some kinases able to phosphorylate serine or threonine residues (serine-threonine kinases) and some able to phosphorylate tyrosine residues (tyrosine kinases). However, they can also be divided into a number of smaller groups based on their substrate specificities, their biological roles, or the nature of their structural and catalytic domains [Hanks and Hunter, 1995, Miranda-Saavedra and Barton, 2007]. For instance, the CMGC family of kinases includes some responsible for regulating the cell cycle, as well as metabolic and stress responses. An example of a specific member of this group is CDK2, which—in addition to containing several phosphorylation sites, as mentioned above—is itself a protein kinase and thus catalyzes the phosphorylation of other proteins. A molecular model of CDK2 is shown in Figure 2.5.

2.1.6 Cellular signaling pathways

A cellular signaling pathway is a series of chemical reactions and other biological events that bridge the gap between the detection of a stimulus and the change in cellular physiology that occurs as a response. The phosphorylation or dephosphorylation of proteins is a common mechanism for transmitting these signals. For instance, insulin is a hormone that regulates the production and breakdown of energy-storing carbohydrates. However, it does not perform this regulation directly; instead, when insulin interacts with a receptor on the surface of a cell, a cascade of chemical reactions occur, ultimately resulting in the conversion of glucose to glycogen, which is a large carbohydrate used as long-term energy storage. Several of the reactions in this pathway involve the phosphorylation of proteins. For instance, glycogen is synthesized by the enzyme glycogen synthase, which becomes inactivated when phosphorylated. Glycogen synthase is phosphorylated by another enzyme called glycogen synthase kinase (GSK), which also becomes inactivated when phosphorylated. Thus, one of the events initiated by insulin (via other signaling events not described here) is the inactivation of glycogen synthase kinase by yet another kinase called AKT (not an acronym), which is sometimes called protein kinase B [van Weeren et al., 1998]. Figure 2.6 illustrates this pathway.

Signaling often occurs via cascades, with one protein catalyzing the phosphorylation of a second protein, which in turn catalyzes the phosphorylation of a third protein, and so on. Because the number of proteins re-

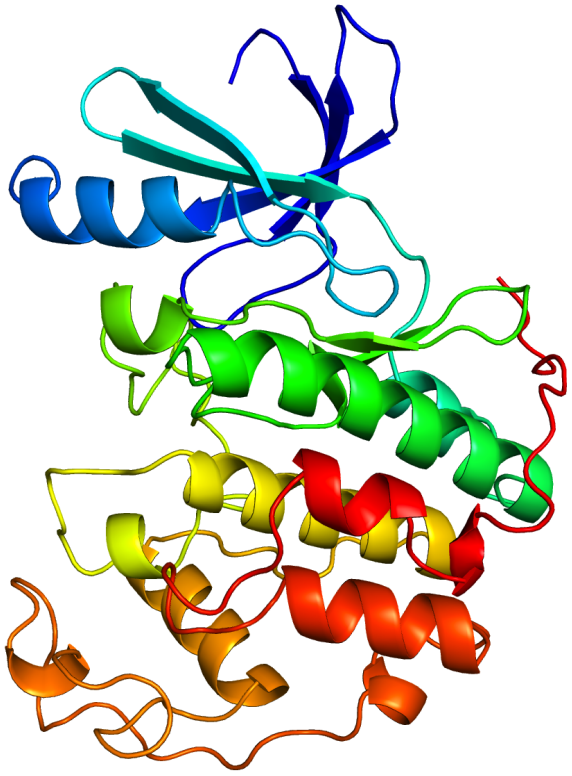


Figure 2.5: Cartoon representation of the three-dimensional structure of cyclin-dependent kinase 2 (CDK2). Ribbons represent a type of local structure called an α-helix, while arrows represent another type of local structure called a β-sheet. Narrow cylinders represent loops or disordered regions. This figure was created using Pymol [Schrödinger, LLC, 2010] from the Protein Data Bank [Berman et al., 2000] entry 1HCL [Schulze-Gahmen et al., 1996].

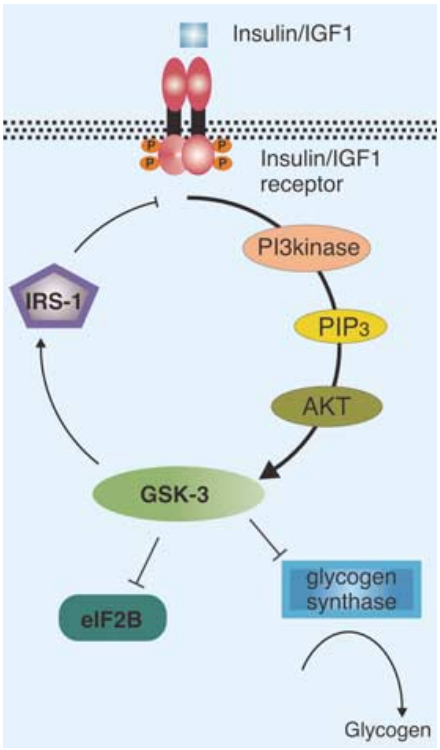


Figure 2.6: An illustration of the pathway by which the release of insulin results in the synthesis of glycogen. This figure was reproduced with permission from a paper by Gould and Manji [2005].

ceiving a signal usually multiplies in each step in the cascade, this allows a weak initial signal to ultimately be amplified into a very strong signal. For example, mitogen-activated protein kinases (MAPK) regulate many cellular processes, including cell proliferation, cell differentiation, and responses to environmental stress [Pearson et al., 2001]. MAPKs can be regulated by a mitogen-activated protein kinase kinase (MAPKK), a type of protein that catalyzes the phosphorylation of MAPKs; similarly, MAPKKKs use phosphorylation to control the activity of MAPKKs [Chang and Karin, 2001]. The level of activity of a given signaling pathway is constantly being regulated in response to changes in environmental conditions. When a stimulus causes an increase in the activity of a given pathway, that pathway is said to be upregulated; similarly, a pathway is downregulated when a stimulus causes a decrease in its activity.

2.2 Kinome microarrays

In the past, the study of biochemistry and cell biology was dominated by the use of low-throughput techniques—those allowing the analysis of only one (or at most a few) biomolecules at a time. While such techniques are useful and informative, they can be time-consuming and tedious to perform, and are often not conducive to understanding a biological system as a whole. The incredible complexity of biological systems demands high-throughput techniques in order to be able to understand them from a global perspective.

One of the most prominent examples of a high-throughput biological technique is the microarray. Microarrays consist of a two-dimensional array of biomolecules (such as DNA [Ramsay, 1998], proteins [Lueking et al., 1999], peptides [Houseman et al., 2002, Houseman and Mrksich, 2002], or carbohydrates [Dyukova et al., 2006]) affixed to a solid support (often a glass slide). A single array may contain hundreds or even thousands of different biomolecules. Once it has been fabricated, a microarray is used by incubating it with a biological sample, and then testing for some sort of interaction between the molecules in the sample and those on the array.

The most common type of microarray is the DNA microarray, which is typically used to study how gene expression (see Section 2.1.1) changes under different conditions—say, in cancerous cells versus non-cancerous cells, or in cells exposed to a particular drug versus non-exposed cells [Schena et al., 1995]. DNA microarrays facilitate this analysis by allowing the amount of each mRNA present in the cells to be measured. Specifically, each gene in the organism being studied is represented on the array by a short DNA sequence that is a subsequence of that gene. Each "spot" on the array contains many copies of one of these sequences. When the mRNA from a cell extract is reverse-transcribed to cDNA and then exposed to the array, a given cDNA may be complementary to one of the DNA molecules on the slide and will thus interact ("hybridize") with it. The amount of cDNA hybridized to each spot can be measured, which indicates the level of expression of the gene corresponding to that sequence.

Whereas DNA microarrays are composed of short sequences of DNA, kinome microarrays consist of short peptides (approximately 15 amino acids in length) that act as substrates for phosphorylation by protein

kinases [Houseman et al., 2002, Houseman and Mrksich, 2002]. As phosphorylation plays an integral role in cellular signaling processes (see Sections 2.1.4–2.1.6), kinome microarrays can be used to study how these processes change in response to different conditions. The following sections discuss kinome microarrays in detail. Specifically, Section 2.2.1 gives a general description of kinome microarrays, while the design of kinome arrays is discussed in Section 2.2.2. Section 2.2.3 summarizes the procedure for using kinome microarrays in the laboratory, while the design of experiments using kinome arrays is covered in Section 2.2.4. Finally, Section 2.2.5 reviews previous studies describing the use of kinome microarrays to provide insight into biological systems.

2.2.1 General description of kinome microarrays

Kinome microarrays are small, rectangular glass slides (approximately 7.6 cm long and 2.5 cm wide) containing hundreds or thousands of spots arranged in a grid-like pattern. Each spot contains many copies of a short peptide of a particular sequence. A picture of a kinome microarray after incubation with a biological sample is shown in Figure 2.7. The exact layout of a kinome array may differ depending on factors like the number of unique peptides, the number of intra-array technical replicates per unique peptide, and the technology used for placing the peptides on the slide. On this particular array, the spots are printed in several "levels" of organization. Specifically, the array contains three level A blocks, each of which contains four level B blocks, each of which contains three level C blocks. The relevance of the colour of each spot will be discussed in Section 2.2.3.

The sequence of each peptide on a kinome microarray is a subsequence of a full-length protein in the organism being studied, and contains a residue that is known or suspected to be phosphorylated in that protein. For example, consider the human protein cyclin-dependent kinase 1 (CDK1), which contains numerous experimentally characterized phosphorylation sites. The sequence of CDK1 is shown in Figure 2.8, with its known phosphorylation sites highlighted in blue. Given this protein sequence, it is easy to derive 15-mer peptides (the length often used for kinome microarrays) containing phosphorylation sites as their central residues. For instance, the peptides corresponding to the phosphorylation sites S46, Y77, and T141 would be `SEEEGVPSTAIREIS`, `LMQDSRLYLIFEFLS`, and `LLIDDKGTIKLADFG`, respectively. In some cases, a phosphorylation site can be too close to the N-terminus or C-terminus of the protein to create a peptide with a full 7 residues on either side. For instance, there are only three residues on the N-terminal side of the phosphorylation site Y4. Two options for dealing with this problem naturally present themselves: use a shorter peptide (`MEDYTKIEKIG`), or include more residues on the C-terminal side in order to make a peptide that is the full 15 residues in length (`MEDYTKIEKIGEGTY`).

2.2.2 Designing kinome microarrays

Designing a kinome microarray refers to the process of selecting peptides to include on the array. Ideally, each peptide chosen should have an experimentally-identified phosphorylation site as its central residue.

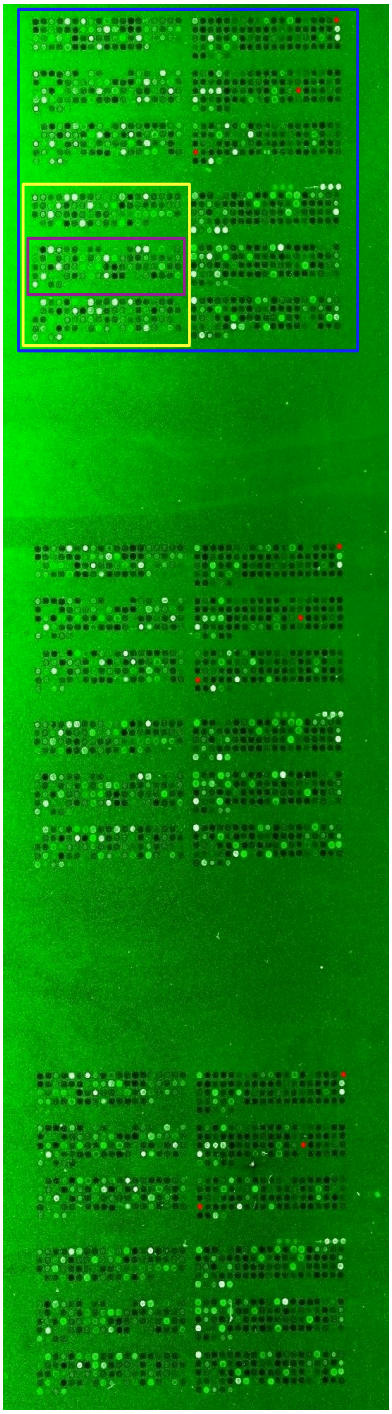


Figure 2.7: Scanned image of a kinome microarray after incubation with cell lysate and staining. Black spots contain peptides that underwent little or no phosphorylation, green spots represent moderate amounts of phosphorylation, and white spots denote a high degree of phosphorylation. As described in the text, the layout of the spots comprises three levels of organization. The blue box surrounds the first level A block; the yellow box surrounds the third level B block within the first level A block; the purple box surrounds the second level C block within the third level B block. Nine of the spots were coloured red using an image-editing program; these represent the nine intra-array technical replicates corresponding to a single peptide sequence.

```
  1  MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRH
 61  PNIVSLQDVLMQDSRLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCH
121  SRRVLHRDLKPQNLLIDDKGTIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSAR
181  YSTPVDIWSIGTIFAELATKKPLFHGDSEIDQLFRIFRALGTPNNEVWPEVESLQDYKNT
241  FPKWKPGSLASHVKNLDENGLDLLSKMLIYDPAKRISGKMALNHPYFNDLDNQIKKM
```

Figure 2.8: Sequence of the human protein cyclin-dependent kinase 1 (CDK1). Residues that have been experimentally determined to be phosphorylation sites according to the PhosphoSitePlus database [Hornbeck et al., 2004, 2012] are coloured in blue. The sequence is shown with 60 amino acids per row; the position number of the first residue in each row is indicated to the left of that residue. The sequence shown corresponds to the UniProt [Apweiler et al., 2004, Boutet et al., 2007, UniProt Consortium, 2008, 2013] accession number P06493.

Although such peptides can be identified by searching the literature, this would be extremely time-consuming. Fortunately, several online, curated databases of literature-derived phosphorylation sites exist. Two such databases—PhosphoSitePlus [Hornbeck et al., 2004, 2012] and Phospho.ELM [Diella et al., 2004, 2008, Dinkel et al., 2011]—contain phosphorylation sites from a wide variety of organisms. Specifically, PhosphoSitePlus contains thousands of sites from human, rat, and mouse, a few hundred sites from chicken, cow, rabbit, and pig, and a small number of sites from other organisms. Phospho.ELM contains far fewer sites than PhosphoSitePlus for the organisms mentioned above, but does contain a substantial number of sites for two organisms poorly represented in PhosphoSitePlus: the fruit fly *Drosophila melanogaster*, and the nematode *Caenorhabditis elegans*. More specialized databases also exist; for instance, P^3DB [Gao et al., 2009b, Yao et al., 2012] contains only phosphorylation sites from plants, while PhosphoGRID [Stark et al., 2010, Sadowski et al., 2013] is limited to yeast. In addition to the protein and the location of the phosphorylated residue within that protein, each record in a given database lists the publication(s) that reported the discovery of the phosphorylation site. If available, information regarding the kinase that catalyzes the phosphorylation of a particular site may also be given. However, most phosphorylation sites are now discovered using mass spectrometry, which does not give information regarding the kinase associated with each site; therefore, most records in the phosphorylation site databases do not contain this information. For instance, all of the data in P^3DB are derived from studies using mass spectrometry, and thus none of its records contain kinase information. PhosphoSitePlus maintains two separate databases—one for records containing kinase information, and one for records that do not. The latter database currently contains 256,877 records, while the former database contains just 13,903.

While it is preferable to choose experimentally-determined phosphorylation sites, this may not be possible if few sites from the organism of interest have been identified. For instance, there are currently 160,753 human sites in the PhosphoSitePlus database, so there are many sites from which to choose when designing a kinome microarray for studying human. Conversely, the same database contains only 12 sites from sheep,

so the designer of a kinome array for studying sheep would not be able to rely solely (or even primarily) on experimentally-characterized phosphorylation sites.

2.2.3 Obtaining and using kinome microarrays

Once the peptides for a kinome microarray have been chosen, the arrays themselves must be fabricated. As few laboratories have the equipment necessary to synthesize peptides and spot them on a slide, kinome arrays are usually obtained from commercial providers. After the arrays have been obtained, they can be used to characterize the protein kinase activity in the system of interest. While a detailed explanation of the laboratory procedure for using kinome arrays is beyond the scope of this thesis, the process will be summarized briefly here. First, the cells being studied are broken open (lysed), and the array is exposed to the contents of the cells (incubated) for approximately 2 hours. During this time, the protein kinases from the cells will catalyze the phosphorylation of the peptides on the array. The number of peptides of a particular sequence that become phosphorylated will depend on the number of protein kinases in the cell lysate that catalyze its phosphorylation, as well as the level of activity of those kinases. Next, the array is exposed to a fluorescent stain that binds to peptides that are phosphorylated, but not to peptides that are not phosphorylated. When the stain absorbs light at a certain wavelength, light is emitted at another wavelength. The intensity of the emitted light can be measured using an image scanner, which indicates the amount of stain bound to a particular spot, which in turn indicates the degree to which the peptides on that spot were phosphorylated. In Figure 2.7, spots having little or no phosphorylation are black, because the lack of bound dye means little light gets emitted. Spots having moderate phosphorylation are green, because the emission wavelength of the dye is in the green range. Finally, heavily phosphorylated spots are white due to the high intensity of the emitted light. Because the values read by the image scanner are related to the intensity of the emitted light, they are often called "intensity values" or "intensity measurements". Readers interested in a more detailed description of laboratory aspects of kinome microarrays, including specific reagents and experimental conditions, can consult Jalal et al. [2009].

Because the stain can also bind non-specifically to the slide itself, the intensity value near the spot (the local background intensity) is subtracted from the intensity value of the spot itself (the foreground intensity), the result of which indicates the level of intensity resulting specifically from the phosphorylated peptides in that spot. Local background intensities are used (rather than, say, the average background intensity over the entire slide) because the background intensity often varies in different parts of the slide.

Once measurements have been obtained from at least two kinome arrays, these data can then be subjected to further analysis and biological interpretation. Kinome array data can be analyzed in terms of the differential phosphorylation of individual peptides (see Section 2.3.4) and the differential modulation of signaling pathways (Section 2.3.5). Samples can also be analyzed by comparing the kinome profile of each, which is defined as the combined phosphorylation intensities of all of the peptides on the array. This is often done using clustering, which is covered in Section 2.3.6.

2.2.4 Microarray experiment design

This section discusses basic principles behind designing statistically and biologically valid microarray experiments, including the concepts of treatment and control arrays, technical and biological replicates, and biological subtraction.

Treatment and control arrays

Because it is difficult to measure absolute levels of phosphorylation using kinome microarrays, the phosphorylation levels measured from a single array rarely have meaning by themselves; they only acquire meaning when compared to the phosphorylation levels from other arrays. Thus, all microarray experiments involve, at a minimum, two arrays—a treatment array and a control array. A treatment array is one in which the biological sample is of interest, and a control array provides baseline measurements to which the treatment array can be compared. For example, a control array might be prepared using a biological sample taken from a healthy animal, while a treatment array might be prepared using a sample from an animal having some bacterial infection. Of course, it is possible to have more than one treatment. For instance, another array could be incubated with a sample from an animal that was infected with the same bacterium, but naturally exhibits no symptoms; yet another sample could be taken from an infected animal that is treated with an experimental drug. The concept of treatment and control arrays also exists in time-course experiments, in which the control sample is the one taken before the treatment is administered.

The use of treatment and control arrays makes it possible to determine that a given peptide exhibits, say, three times the amount of phosphorylation in a treatment array compared to the control array, in which case it is likely that the treatment has an effect on the phosphorylation of that peptide. Such a peptide would be termed "differentially phosphorylated". A major goal of kinome microarray experiments is to determine how the treatment arrays differ from the control array (and perhaps also from each other) in terms of the phosphorylation level of each peptide. Once these are known, differences in the modulation of entire biological pathways can be detected. Procedures for doing this are described in Chapter 7.

Technical replicates

All biological experiments are subject to the effects of random variation. In order to evaluate the degree to which random error affects a particular measurement, that measurement must be performed multiple times in exactly the same way. These repeated measurements are called technical replicates. In the context of kinome microarrays, there are two methods by which technical replicates can be performed. First, multiple, identical arrays can be incubated with the same biological sample (inter-array replicates). Second, each peptide of a particular sequence can be spotted multiple times on the same array (intra-array replicates), and the measurements for corresponding spots (i.e., those containing the same peptide) can be compared. In practice, the latter method is typically used. In Figure 2.7, the nine technical replicates corresponding to a

single unique peptide sequence are coloured in red.

Performing multiple technical replicates has three benefits. First, the phosphorylation measurements for a given peptide can be pooled and averaged, reducing the effect of random variation. For example, if three technical replicates are performed, and the phosphorylation measurements for a particular peptide are (in arbitrary units) 5.2, 7.3, and 6.1, then the three measurements can be averaged together to get a measurement of 6.2, which can be considered more reliable than any of the three individual measurements. Second, peptides whose measurements exhibit an abnormal amount of variation among the technical replicates can be identified. For example, if the three measurements for a given peptide are 0.5, 9.8, and 28, then—assuming that the scale of the measurements is such that these numbers represent an unusually large degree of variation—that peptide could be excluded from subsequent analyses. Third, multiple technical replicates are necessary for assessing statistically whether the level of phosphorylation of a given peptide in one sample is different than that in a second sample (see also Section 2.3.4).

Biological replicates

Whereas technical replicates control for random variation, biological replicates control for biological variation. For instance, suppose that a researcher is examining how signaling patterns differ in healthy pigs versus those infected with some virus. Furthermore, suppose that one sample is taken from a healthy pig and one sample is taken from an infected pig, and the resultant kinome microarray data indicate the presence of substantial differences in a particular signaling pathway between the two samples. Given just these two samples, it would not be valid to attribute these differences to the presence or absence of the virus, since the differences could just as easily reflect other environmental or genetic characteristics of the two subjects. Thus, multiple biological replicates—in other words, samples from many infected and many non-infected individuals—would be required in order to determine the true effect of the virus on cellular signaling patterns.

Biological subtraction

Unfortunately, using multiple biological replicates is sometimes insufficient to identify the true effect of a treatment on the system in question. This is because the genetic background of a biological subject can have a substantial influence on its kinome profile—so substantial that it can partially or even completely mask the effect of the experimental treatment. For instance, suppose that the effect of a certain virus on the kinome profile of outbred (i.e., genetically distinct) pigs is being investigated. Further suppose that 10 pigs are infected with the virus and 10 different pigs are not, and that kinome microarray analysis is performed on cell extracts taken from all 20 pigs. If the resulting kinome profiles were clustered, the pigs may not cluster by infection status, as one might expect. Instead, the genetic background of the pigs may have a greater influence on the clustering pattern, with genetically similar pigs clustering together.

To circumvent this problem, the experiment could be redesigned as follows. Cell samples would first be taken from all 20 pigs. Subsequently, 10 of the pigs would be infected with the virus, while the other 10

pigs would be subjected to a "mock infection", which would involve the same intervention (e.g., injection via a syringe) as in the first group, except lacking the actual virus. After a period of time, cell samples would again be taken from all 20 pigs, followed by kinome microarray analysis of those samples. Prior to clustering, the intensity values of each peptide from each control sample (i.e., those taken prior to infection or mock-infection) would be subtracted from the value for the same peptide in the infected or mock-infected sample from the same animal ("biological subtraction"). Theoretically, since this subtraction removes the effect of the genetic background of each subject, the resultant values should reflect only the impact of the intervention (infection or mock-infection) on the kinome profiles of the subjects. In this case, it is more likely that the infected pigs and the mock-infected pigs would segregate from one another in the clustering analysis.

Unfortunately, this type of experimental design is not always possible. This can be due to either ethical or practical reasons. To illustrate the former category, suppose that humans were being studied instead of pigs in the above example. If the virus has the potential to cause serious illness or death, then it would be unethical to purposely infect the subjects of the study with the virus. One alternative would be to take samples from a large number of healthy individuals, and then wait for a period of time. During that time, some of the individuals will naturally contract the virus, and samples can then be taken from them. However, such a study may be both time-consuming (it may take a long time for a sufficient number of people to contract the virus) and expensive (a large number of control samples must be taken so that a sufficient number of individuals are likely to eventually contract the virus). Thus, a faster and cheaper (but less effective) study design is simply to compare samples from individuals with the virus to samples from individuals without the virus. However, this has the disadvantage that the genetic background of the individuals may have a greater effect on the results than the presence or absence of the virus.

As an example of a practical reason preventing the use of biological subtraction, consider a study involving honeybees. Because of their fragility and small size, it may be difficult to obtain a cell sample from a honeybee without destroying the animal, which abrogates the possibility of obtaining a sample from the same organism both before and after the administration of an experimental treatment.

Biological subtraction also plays a role in time-course experiments, in which samples are taken from the same animal both before any intervention is made (time 0) and at various time points post-intervention. In this kind of experimental design, biological subtraction can be performed by separately subtracting the time 0 kinome profile of a given animal from the kinome profile corresponding to each post-intervention time point.

2.2.5 Studies applying kinome microarrays to biological problems

This section reviews existing literature describing the use of kinome microarrays to investigate biological systems. The first such study was published by Diks et al. [2004]. They first tested the validity of their array by incubating it with the catalytic portion of protein kinase A (PKA), and found that—as expected—peptides containing PKA recognition sequences were phosphorylated, while those not containing such sequences re-

mained unphosphorylated. The array was then used to study changes in phosphorylation patterns in response to stimulation by lipopolysaccharide (LPS), a component of the cell wall in certain bacteria that activates a number of biological pathways related to the immune system. The authors observed that the data derived from their kinome microarray experiments were consistent with those derived from other biological techniques used to study LPS, giving further confidence that the array data were valid and biologically meaningful. The authors also discovered that a protein called p21Ras is activated by LPS, a novel result confirmed by the authors using further experiments.

Following the publication just described, a number of additional papers applying kinome microarrays to biological problems have been published. A selection of these papers is summarized below; to emphasize the wide variety of biological problems and systems that kinome microarrays can be used to investigate, these papers are organized by the biological problem being addressed.

Cancer

In one study applying kinome microarrays to cancer, van Baal et al. [2006] compared the kinase-associated signaling in biopsies of Barrett's Esophagus, an esophageal lesion that sometimes progresses to cancer, with biopsies of two adjacent tissue types. The authors discovered that the signaling profile of Barrett's Esophagus was intermediate between those of the two adjacent tissues. They also searched for signaling pathways that were upregulated or downregulated in Barrett's Esophagus, and found that enzymes related to glucose metabolism were upregulated.

In another study, Schrage et al. [2009] examined signaling patterns in chondrosarcoma, a cancer of cartilage-producing cells. After identifying signaling pathways that were upregulated in the cancerous cells, the authors tested two different drugs that have been shown to inhibit these pathways. One of these drugs, called dasatinib, resulted in significantly reduced growth of the cancer cells, suggesting that it might be a useful treatment for chondrosarcoma.

Immunosuppressants

In 2005, Löwenberg et al. used a kinome microarray to characterize the effect of dexamethasone (a member of the glucocorticoid family of drugs, which are commonly prescribed as immunosuppressants) on cellular signaling pathways, and found that substrates of the protein kinases Lck and Fyn had significantly lower levels of phosphorylation in cells treated with dexamethasone compared to control cells.

In a related study, kinome microarrays were used to examine the molecular basis for a serious side effect of glucocorticoid treatment—insulin resistance [Löwenberg et al., 2006]. More specifically, the authors examined the short-term (i.e., less than 30 minutes after administration) effect of dexamethasone on the kinome profile of adipocytes, a type of cell linked to insulin resistance. Their analyses revealed that dexamethasone-treated adipocytes exhibited reduced phosphorylation of targets of the insulin receptor kinase, as well as reduced activities of several downstream kinases in the insulin signaling pathway. Perhaps most interestingly, the

authors showed that these changes in cellular signaling were not due to changes in transcription. Their work has implications in the search for glucocorticoids that retain the immunosuppressive properties associated with this class of drugs, but lack the ability to induce insulin resistance.

Kidney disease

de Borst et al. [2007] used kinome microarrays to compare protein kinase activities in three different groups of rats. The first group was comprised of Ren2 rats, a rat model exhibiting kidney disease caused by defective angiotensin II, a short peptide involved in the regulation of blood vessels. The second group consisted of Ren2 rats treated with angiotensin-converting enzyme inhibitor, a protein known to reverse the effects of angiotensin II-mediated renal damage. As a control, the third group contained rats not associated with kidney disease. The authors found that a number of kinase targets were differentially regulated in the Ren2 rats compared to the control rats, and these changes were partially or totally abrogated in the Ren2 rats treated with angiotensin-converting enzyme inhibitor.

Infectious diseases

In one study that applied kinome microarrays to the study of infectious disease, Kindrachuk et al. [2012] compared how cells respond to infection by two varieties of monkeypox virus (MPXV): Congo Basin MPXV and West African MPXV. While these viruses are similar, the former is lethal in about 10% of cases, whereas the latter is less virulent and rarely deadly. As might be expected, the authors found that certain pathways related to the immune system were significantly downregulated by Congo Basin MPXV but not by West African MPXV, an observation that likely reflects the greater virulence of Congo Basin MPXV.

In another study, Arsenault et al. [2013b] examined the cellular response in chickens to *Salmonella enterica* serovar Typhimurium, which is a zoonotic bacterium that can cause severe symptoms in humans, but causes little or no pathology in chickens when infected more than a day after birth. To examine the cellular response to infection over time, this study involved infecting or not infecting five-day-old chickens with the bacterium, and then sacrificing the birds 1, 4, 7, or 21 days post-infection so that muscle samples could be extracted. A number of differences were observed in the infected birds compared to the uninfected ones, particularly in pathways related to carbohydrate metabolism. Whereas most studies (whether using kinome microarrays or not) describe differences observed in symptomatic infections, this study is of interest in that it reveals that several cellular changes take place despite the infection not resulting in obvious pathology.

Plant signaling

Kinome microarrays are not limited to studies on animals; they can also be used to examine cellular signaling in plants. For example, following an exploratory study that confirmed the applicability of kinome arrays to samples from plants [Ritsema et al., 2007], the same authors used the technology to investigate the roles of jasmonic acid and salicylic acid—hormones that are known to induce defences against plant pathogens—on

kinome responses in *Arabidopsis thaliana* [Ritsema et al., 2010]. In two additional studies, the same research group investigated signaling pathways associated with sugar metabolism in plants [Ritsema et al., 2009, Ritsema and Peppelenbosch, 2009]. Interestingly, all of these studies used arrays containing peptides from many different organisms—not just from plants. Justification for this can be found in another study by these authors, which argues that when cell extracts from different organisms are applied to the same array, they exhibit similar phosphorylation patterns, suggesting the existence of a "minimal eukaryotic phosphoproteome" [Diks et al., 2007]. The strategy of using the same kinome array to study different species differs from the one used in this thesis, which focuses in part on the design of species-specific arrays (Chapters 3-6).

2.3 Computer science concepts

This section gives the computer science background necessary to understand the remainder of this thesis. Specifically, Section 2.3.1 describes BLAST, a widely-used tool for searching databases of DNA or protein sequences. Section 2.3.2 discusses concepts relating to classification problems and machine learning, while Section 2.3.3 explains why microarray data must be preprocessed before being analyzed, as well as techniques for doing this. Statistical tests for identifying differentially phosphorylated peptides in kinome array data are described in Section 2.3.4, while the identification of differentially modulated signaling pathways is discussed in Section 2.3.5. Finally, clustering techniques are covered in Section 2.3.6.

2.3.1 BLAST

A common task in bioinformatics is as follows: given a nucleic acid or protein sequence, identify similar sequences in a database. One method that could be used to do this is to perform an alignment between the query sequence and each of the database sequences. The Smith-Waterman dynamic programming algorithm can be used to find an optimal local alignment between two sequences of length n and m in $\mathrm{O}(nm)$ time [Smith and Waterman, 1981]. A local alignment is one that does not necessarily involve all of the letters in each sequence. Assume for simplicity that all of the sequences in the database are of length m, and that there are d sequences in the database. Then the time required to search the database would be $O(nmd)$. Given the large size of many sequence databases (for example, the National Center for Biotechnology Information (NCBI) protein database contained 35,616,906 sequences as of January 2014), this would be computationally impractical. Thus, heuristic methods are required in order to search sequence databases more quickly.

The Basic Local Alignment Search Tool (BLAST) [Altschul et al., 1990, 1997] is a popular heuristic-based method for searching databases of protein or DNA sequences. Several variants of BLAST exist; which one is used in a particular situation depends on the nature of the query sequence and database. For example, `blastn` is used to search a nucleotide query against a nucleotide database, and `blastp` is used to search a protein query against a protein database. The BLAST variant that is of interest in this thesis is `blastp`.

The algorithm for `blastp` works as follows. First, the query sequence is split into words of length three

(although the word length can be changed by the user). For example, the query sequence QGFTPETRK would be split into the words QGF, GFT, FTP, TPE, PET, ETR, and TRK. For each word, a list of similar words is then compiled. "Similar" is defined in terms of an amino acid substitution matrix, which contains scores for substituting one amino acid with another. A commonly-used substitution matrix is BLOSUM62 [Henikoff and Henikoff, 1992], which is shown in Table 2.2. Substitution matrices assign high scores when an amino acid does not change, or when an amino acid is replaced by another with similar chemical properties. For instance, Table 2.2 shows that substituting Arg (see Table 2.1 for a mapping between three-letter codes and one-letter codes) for itself gives a score of 5, while substituting Arg with Lys—both of which are positively-charged amino acids—has the smaller, but still positive, score of 2. However, substituting Arg with Asp (a negatively-charged amino acid) has a score of -2. Words similar to a given word in the query sequence are identified by using the substitution matrix to compute the score for the query word against all possible three-letter words. The score for a given word is simply the sum of the substitution scores for all three positions. For instance, the score for TPE compared to TPD is $5+7+2=14$ when using BLOSUM62, while the score for TPE compared to VYW is $0+(-3)+(-3)=-6$.

Words with a score higher than a threshold T are used in the next stage of the algorithm. An appropriate value of T depends on the scoring matrix used. In the next stage, a finite-state automaton (FSA) data structure [Hopcroft et al., 2006] is created from all of the high-scoring words. When a database sequence is run through this automaton, it will end in a final state if and only if it contains one of the high-scoring words. For each such database sequence, the high-scoring word that was found within it is used as a "seed" for an alignment between that sequence and the query sequence. Starting with the seed, the alignment is extended outward in both directions. As each new position is added, the current score S of the alignment is calculated from the substitution matrix. The highest score H obtained thus far is also retained. The extension continues until S drops off by a certain amount from H, or until the end of one of the sequences is reached. The alignment returned by BLAST is the one that produced the highest score (not necessarily the longest alignment). If two neighbouring alignments were found in the same database sequence, they are then combined into a longer alignment.

For example, suppose that the sequence mentioned above (QGFTPETRK) was used as a BLAST query against a protein database containing the sequence PGYTPDTRC, and that the word TPD (which is contained in the database sequence and is, as shown above, similar to the word TPE found in the query) is used as the seed. The process of forming an alignment by extending this seed is illustrated in Figure 2.9. As the highest score is found after the second extension, the corresponding alignment is the one that would be returned by BLAST. Note that for brevity, this illustration shows the extension as if it occurs simultaneously on both the left and right sides of the seed. However, the extension of the seed to the left and right would actually occur independently; for instance, a seed could be extended just a few residues to the left but many residues to the right.

The statistical significance of a match between a query sequence and a database sequence is given in

Table 2.2: The BLOSUM62 substitution matrix. As the matrix is symmetric, only the lower triangle is shown. Column and row headings are one-letter amino acid codes, and the values represent the score associated with substituting the row amino acid with the column amino acid or vice versa.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 2.9: The creation of an alignment by BLAST between the query sequence `QGFTPETRK` and the database sequence `PGYTPDTRC` using the word `TPD` as the seed. Scores are calculated using the BLOSUM62 substitution matrix (see Table 2.2). A vertical bar between two residues in the alignment indicates an exact match, a colon indicates a conservative substitution (one with a BLOSUM62 value greater than 0), and a period indicates a non-conservative substitution.

the form of an "E-value", which is defined as the expected number of matches having a score equal to or greater than H that might occur by chance given the size of the database. While the E-value threshold for considering a match to be significant can vary depending on the application, 10^{-3} is a commonly-used value.

A formal analysis of the computational complexity of BLAST would be fairly complicated; however, the extension of seeds accounts for the majority of its running time [Altschul et al., 1997]. Although run times will vary depending on the computing power used, as well as the nature of the query and the database, a single query can usually be searched against a large database in less than a minute.

Several implementations of BLAST are available. The most common one is provided by NCBI, and is available both as a web-based tool (`http://blast.ncbi.nlm.nih.gov`) and as a stand-alone program that can be run on the user's local machine. The web-based tool is ideal for searching a single sequence (although it does have a limited batch mode), while the stand-alone tool is useful for searching hundreds or thousands of sequences, as well as for searching custom databases (a feature not offered by the web-based tool). Another implementation is Washington University BLAST (WU-BLAST; `http://www.ebi.ac.uk/Tools/sss/wublast`), which has a modified interface and different databases compared to NCBI BLAST.

2.3.2 Classification problems and machine-learning classifiers

Classification problems are those in which entities must be placed into the correct category. For instance, geologists classify rocks as igneous, sedimentary, or metamorphic, while words can be classified according to their part of speech (noun, verb, etc.). A single entity in a given problem is called an instance; for example, a specific rock would be an instance in the former problem, while the word "hello" would be an instance in the latter problem. Performing a correct classification requires taking into account the attributes of the instance being classified: the density, colour, shape, or hardness of a rock might help a geologist identify the correct category, while the fact that a word ends in "ly" might suggest that it is an adverb (although not always—consider "family").

A machine-learning classifier is a computer program that attempts to categorize instances based on their attributes (also called "features"). A classifier is described as supervised if instances having known classifications are used to train it. A supervised classifier is a function whose input consists of the features of a particular instance and whose output is the category to which that instance is predicted to belong. A classifier is unsupervised if it attempts to find general structure in the data without the benefit of already-known classifications. This section discusses supervised classifiers, while clustering—a type of unsupervised classification—is discussed in Section 2.3.6. Also, this section focuses exclusively on binary classification problems—that is, problems for which there are only two classes. The classes in a binary classification problem can usually be characterized as either positive (e.g., a cancerous tissue sample) or negative (e.g., a non-cancerous tissue sample).

Types of classifiers

Many different types of functions can be used for a classifier, ranging from trivial to complex. One of the simplest possible classifiers is one that always predicts the category that was most common in its training data. For instance, if this classifier was categorizing tissue samples and was trained using 100 cancerous samples and 200 non-cancerous samples, then all unknown samples would be classified as non-cancerous. Obviously, this strategy would not make for an accurate classifier. An example of a more complex classifier is a decision tree, which is a tree where leaves represent classes and internal nodes represent decisions [Quinlan, 1992]. Each internal node tests some attribute of the instance; for instance, a classifier attempting to identify a piece of fruit might test whether the fruit is yellow—if yes, then banana and apricot would still be possibilities, while orange and cherry would be eliminated from consideration. Internal nodes may lead to leaf nodes—in which case the classification has been made—or to other internal nodes, in which case another attribute of the instance is considered. If the next internal node in the fruit classifier tested whether the fruit was round, then an affirmative answer would further eliminate banana.

Some classifiers are modified or extended versions of other classifiers. A random forest, for example, is a classifier that involves generating many decision trees. If X represents the set of features associated with the instances being classified, each such decision tree is generated each using a different randomly-selected subset of X. The class assigned by the random forest is then the class assigned by the majority of its component decision trees.

In addition to decision trees and random forests, another commonly-used classifier is the support vector machine (SVM) [Cortes and Vapnik, 1995]. Based on the features associated with each instance, the instances in the training set are mapped into a high-dimensional space such that those from the positive class can be separated as best as possible from those in the negative class. Yet another is the artificial neural network (ANN) [Bishop, 1996], which is modelled after biological neurons and the connections among them.

Measuring the performance of a classifier

After building a classifier, it is important to determine how accurate it is. There are several different metrics that can be used to measure the accuracy of a classifier [Fawcett, 2006]. To define these metrics, it is helpful to first define some terms. Let TP ("true positives") represent the number of testing instances whose true classification and predicted classification are both positive. Similarly, let TN ("true negatives") represent the number of testing instances whose true classification and predicted classification are both negative. Finally, let FP ("false positives") represent the number of testing instances predicted as positive whose actual class is negative, and FN ("false negatives") represent the reverse. Several performance measures that can be defined in terms of these numbers are given in Table 2.3. Note that some articles in the literature may use alternate definitions for some of these terms; however, the definitions found in Table 2.3 are used in this thesis.

Some types of classifiers, such as decision trees, output only the name of the predicted class [Fawcett, 2006]. Conversely, some types instead output a score for each instance, with a higher score typically indicating

Table 2.3: Measures for evaluating the performance of classifiers. PPV, positive predictive value; NPV, negative predictive value; MCC, Matthews correlation coefficient.

Name	Definition	Interpretation
Sensitivity	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$	Proportion of instances whose actual classification is positive that were classified correctly
Specificity	$\frac{\mathrm{TN}}{\mathrm{TN+FP}}$	Proportion of instances whose actual classification is negative that were classified correctly
PPV	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$	Proportion of instances having a positive prediction that were classified correctly
NPV	$\frac{\mathrm{TN}}{\mathrm{TN+FN}}$	Proportion of instances having a negative prediction that were classified correctly
Accuracy	$\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$	Proportion of instances classified correctly
MCC	$\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP+FP})(\mathrm{TP+FN})(\mathrm{TN+FP})(\mathrm{TN+FN})}}$	Correlation between actual classes and predicted classes

that it is more likely to be a positive instance. For such classifiers, TP, TN, FP, and FN must be calculated in terms of a particular threshold, where instances with scores higher than the threshold are classified as positive and vice versa. Changing this threshold would change the values of these variables, as well as of the performance metrics given in Table 2.3. Decreasing the threshold, for example, will result in an increase in sensitivity but a decrease in specificity. The choice of threshold depends on the application. For a classifier that attempts to categorize tissue samples as cancerous or non-cancerous, it would be appropriate to choose a relatively low threshold (resulting in high sensitivity but low specificity), since misclassifying a cancerous sample as non-cancerous is likely to be more dangerous that misclassifying a non-cancerous sample as cancerous (although the latter situation could also result in serious consequences, such as unnecessary chemotherapy).

Since the measures listed in Table 2.3 vary depending on the chosen threshold, it is also useful to have a threshold-independent measure of accuracy. One such measure can be derived using a receiver operating character (ROC) curve, in which the y axis represents sensitivity and the x axis represents $1-$specificity. To construct the curve, each unique score given to a testing instance is used as a threshold, and the specificity and sensitivity at that threshold are calculated and plotted. In addition, the points $(0,0)$ and $(1,1)$ are plotted, which respectively represent a threshold higher than the highest score given (giving sensitivity 0 and specificity 1) and a threshold lower than the lowest score given (giving sensitivity 1 and specificity 0). The area under this curve, often denoted $\mathrm{A_{ROC}}$, gives a threshold-independent measure of performance, with a value of 1 indicating perfect discrimination and a value of 0.5 being equivalent to random guessing.

Cross-validation

When testing classifiers, it is critical that the data used for evaluating the classifier are different than the data used to train it. The importance of this can be illustrated by considering another type of trivial classifier—one that simply "remembers" its training data. Specifically, the classifier would store the class corresponding to each exact combination of features in its training data, and when predicting for a "new" instance, the classifier would simply consult its mapping of feature combinations to classes. For an instance with a combination of features not encountered in the training data, a class would be randomly assigned. Such a classifier is of little use, as it can predict accurately only for instances for which the class is already known. However, if this classifier was tested using only its training data, it would perform flawlessly, giving a misleading indication of its accuracy. While this is an extreme example, the performance of any classifier will be exaggerated if it is tested using any of the same data with which it was trained.

Given this, a technique called cross-validation is often used to test a classifier. In this technique, the classifier is trained multiple times. Each time, a different portion of the data is used for testing, while the remaining data are used for training. Specifically, to perform n-fold classification, the set of data D is split into n equal-sized subsets $D_1, D_2, \ldots, D_n$, each of size $|D|/n$. If $|D|$ is not evenly divisible by n, then the subsets cannot all be of equal size, but should be made as close in size to one another as possible. In each fold, one of the subsets is used for testing, while all of the other subsets are combined and used for training. For example, in the first fold, the instances in $D_1, D_2, \ldots, D_{n-1}$ are used for training, and the instances in D_n are used for testing. In the second fold, the instances in $D_1, D_2, \ldots, D_{n-2}, D_n$ are used for training, and the instances in D_{n-1} are used for testing. This continues until all n subsets have been used for testing. At this point, the overall performance of the classifier can be evaluated, as predictions are available for all elements of D. Therefore, cross-validation allows the available data to be used to the full extent possible (all data are ultimately used for training and testing) while avoiding the problem described above (since data used to train the model are never used to test it).

While any value of n can be chosen in the range $2 \leq n \leq |D|$, common values include 10 (10-fold cross-validation) and $|D|$ (also called leave-one-out cross-validation). The choice of n can be influenced by the size of the testing dataset and the speed by which the classifier can be trained; if $|D|$ is very large, or the training of the classifier is slow, it might not be practical to choose $n = |D|$ or other large values of n.

Overfitting

When building a classifier, it is important that the classification function describe real relationships between the features of a given instance and the class of that instance. However, if the model contains many features that are not relevant to predicting the correct class, or if the feature values for the training instances contain a lot of noise, then overfitting may occur. This results in the model being able to make accurate predictions for instances on which it was trained, but not for as-yet unseen instances. Thus, choosing features that are informative for predicting the correct class, and ensuring that the values of features for the training set are

as accurate as possible, are important for creating an accurate classifier.

Implementations of classifiers

There are many software programs that implement classification algorithms. Some implement only a single classifier, such as SVM^{light} [Joachims, 1998], while others implement a wide variety of classifiers, such as the machine-learning package Weka [Frank et al., 2004, Witten et al., 2011]. Implementations of machine-learning algorithms are also available as software libraries for a variety of programming languages, enabling them to be used directly within a script or program. For instance, packages are available for the R programming language containing implementations of SVMs and Bayesian networks, among others [Conway and White, 2012].

2.3.3 Preprocessing of kinome microarray data

The raw intensity measurements taken from kinome microarrays must be preprocessed before meaningful comparisons can be made between different arrays. This section discusses normalization, which is the process of bringing the intensity measurements of multiple arrays onto a common scale, as well as variance-versus-mean dependence, which is a phenomenon observed in microarray data where sets of measurements with higher means tend to have higher variances. A method for performing normalization and for alleviating variance-versus-mean dependence is described.

Normalization

When two or more microarray experiments are performed, there will always be at least some systematic error due to small differences in variables such as sample volume, incubation time, and incubation temperature. As a result, the microarrays may have different intensity distributions. To illustrate this, suppose that microarray experiments A and B (representing a treatment and a control) have mean intensity measurements (over all peptides on the array and over all intra-array technical replicates) of 16000 and 8000, respectively. Now consider a comparison between the same peptide on the two arrays, denoted S_A and S_B, with S_A having a mean intensity measurement among the technical replicates of 12000 and S_B having a mean intensity measurement of 6000. Given that the average intensity measurement from array A was twice as high as that from array B, the difference in intensity measurements between S_A and S_B is probably attributable to this systematic bias, rather than to a real difference between the treatment and the control. Systematic biases like these must be eliminated if meaningful comparisons are to be made between the intensity levels of individual peptides on different arrays. The process of eliminating these systematic biases by bringing the intensity values for all of the arrays onto the same scale is called calibration or normalization[1].

[1]The word "normalization" is often used for different purposes in the scientific literature. For example, some authors use it to refer to the process of transforming data so that it has a normal (Gaussian) distribution. "Calibration" is thus probably a better term for describing the process of bringing the intensity values of multiple arrays onto a common scale. However, since "normalization" seems to be more frequently used in literature describing microarray data analysis, this term will be used here.

One normalization method is called centering [Stekel, 2003]. Given n arrays $A_1, \ldots, A_n$ with respective mean intensities $A_1^M, \ldots, A_n^M$, the first step is to divide every measurement in each array by that array's standard deviation. This ensures that each array's intensity measurements have a standard deviation of 1. Second, the array $A_{\max}$ having the highest mean intensity is identified. For each of the other arrays A_i, the value $A_{\max}^M - A_i^M$ is added to each peptide's intensity measurement. This has the effect of equalizing the mean of each array's intensity measurements. Adjusting each mean to that of the array with the highest actual mean ensures that no negative intensities result, since each intensity measurement can only increase. A number of more sophisticated techniques for normalization have also been described [e.g., Kerr et al., 2000, Yang et al., 2002, Tarca et al., 2005, Zhang et al., 2005], one of which is described in detail later in this section.

Variance-versus-mean dependence

A property of microarray experiments that renders data analysis more complicated is that the variance in intensity measurements is usually not consistent for all peptides. Instead, the variance of the measurements typically increases with the mean of those measurements—a phenomenon called variance-versus-mean dependence [Rocke and Durbin, 2001, Huber et al., 2002]. Figure 2.10A contains a scatterplot showing intensity data from a DNA microarray experiment with two technical replicates per unique DNA sequence. The figure shows that the sequences with the highest average intensities also had the greatest amount of variance between the two measurements.

In order to prevent variance-versus-mean dependence from interfering with data analysis, methods are needed that eliminate this dependence. Early studies often assumed that the variance V is related directly (and solely) to the mean μ, without any additive error—in other words, $V = k\mu$ for some constant k [Rocke and Durbin, 2001]. An obvious deficiency with this assumption is that it implies that spots with a mean close to zero have near-zero measurement error, which is unlikely [Rocke and Durbin, 2001]. This problem could easily be rectified by incorporating an additive term representing measurement error that is always present but is unrelated to the mean. However, Huber et al. [2002] cite a number of studies suggesting that variance-versus-mean dependence sometimes deviates significantly from linearity. In some experiments, for example, almost no variance-versus-mean dependence is observed when the mean is small, but a near-linear dependence is noted when the mean is large. This makes simple linear transformations unsuitable for eliminating variance-versus-mean dependence.

VSN: a method for normalization and eliminating variance-versus-mean dependence

VSN, a technique both for eliminating variance-versus-mean dependence and for normalization, was proposed by Huber et al. [2002]. A description of this method is as follows. Because this thesis concentrates on kinome microarrays, the entities on the array will be referred to as "peptides", even though VSN is potentially applicable to any type of microarray.

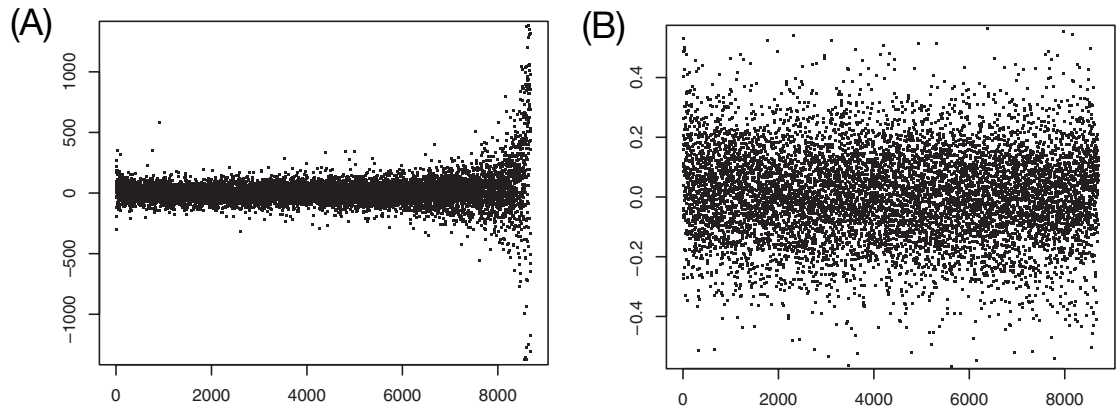


Figure 2.10: The problem of variance-versus-mean dependence. (A) Scatterplot showing results from a DNA microarray experiment performed in duplicate. The x-axis represents the rank of the average of a given spot (with higher ranks having higher averages), and the y-axis represents the intensity measurement from the first replicate minus the intensity measurement from the second replicate. (B) Scatterplot showing the same data as in part (A) after applying the VSN method [Huber et al., 2002]. The x-axis has the same meaning as in part (A), while the y-axis represents the *normalized* intensity measurement of the first replicate minus that of the second replicate. This figure was reproduced by permission from Huber et al. [2002].

Let i represent a particular microarray in an experiment involving d microarrays, and let k represent a particular peptide on those arrays. Also, let y_{ki} represent the raw intensity measurement derived from the image analysis software for peptide k on array i.

The normalization portion of Huber et al.'s model is simple: the authors assume that the intensity values can be normalized using the parameters $o_1, \ldots, o_d$ and $s_1, \ldots, s_d$ such that $\tilde{y}_{ki} = o_i + s_i y_{ki}$, where $\tilde{y}_{ki}$ is the normalized intensity value corresponding to the raw intensity value y_{ki}. This model is later integrated with a model for variance-versus-mean dependence.

Their model for variance-versus-mean dependence is more complex. The authors consider the intensity value for a given spot k to be a random variable Y_k with mean μ_k and variance v_k. Due to variance-versus-mean dependence, v_k is dependent on μ_k, and can be considered a function of it: $v_k = v(\mu_k)$. The authors' objective is to eliminate the dependence of v_k on μ_k, and instead make it have a constant variance σ^2. Based on the work of Rocke and Durbin [2001], the authors assume a quadratic variance-versus-mean dependence function $v(\mu_k)$ involving three parameters. Also, the authors make use of a variance stabilizing transformation $h(y) = \int^y 1/\sqrt{v(u)}\, du$ given by Tibshirani [1988]. The function $h(y)$ has the property that $\mathrm{Var}(h(Y_k))$ is independent of $\mathrm{E}(h(Y_k))$. By inserting $v(\mu_k)$ into the equation for $h(y)$, the authors derive the equation

$$h(y) = \gamma \operatorname{arsinh}(a + by) \tag{2.1}$$

where γ, a, and b are expressions involving the original parameters of $v(\mu_k)$. The function arsinh (also sometimes written arcsinh or $\sinh^{-1}$) is the area hyperbolic sine function, and is related to the logarithm as follows: $\text{arsinh}(x) = \log(x + \sqrt{x^2 + 1})$. The function $h(y)$ is continuous for all y; thus, negative raw intensity values do not pose a problem with Huber et al.'s method.

Finally, the authors combine their model for normalization and their model for variance-versus-mean dependence. Specifically, they replace y in Equation 2.1 with the calibrated intensity value $o_i + s_i y_{ki}$ to get

$$h_i(y_{ki}) = \gamma \, \text{arsinh}(a + b(o_i + s_i y_{ki})) \tag{2.2}$$

where $1 \leq i \leq d$. The statistical model for a peptide k whose intensity level does not change in the different treatments (except for random error) is therefore of the form

$$h_i(Y_{ki}) = \mu_k + \epsilon_{ki} \tag{2.3}$$

where $E(\epsilon_{ki}) = 0$ and $\text{Var}(\epsilon_{ki}) = \sigma^2$ (reflecting the elimination of variance-versus-mean dependence). After transformation, the resulting data have a normal distribution (see also Figure D.2).

In order to estimate the parameters of the above model, Huber et al. use a variant of maximum likelihood estimation. This variant, which is derived in another paper by Huber and coauthors [Huber et al., 2003], estimates the parameters using a set of spots that have similar intensities among the arrays.

The results of applying Huber et al.'s method is shown in Figure 2.10B, which depicts the same data as given in Figure 2.10A, except after applying VSN. As the figure shows, there is no relationship between the rank of the mean and the difference in transformed intensity. The authors compared their method to others, and showed that theirs performed best in eliminating variance-versus-mean dependence.

Of course, eliminating variance-versus-mean dependence is useless in practice if it does not improve the ability to identify differentially expressed genes (or differentially phosphorylated peptides, when applied to kinome microarrays). To provide evidence that their method does, in fact, make a difference in this regard, Huber et al. [2002] used data from DNA microarray experiments involving extracts from both cancerous and non-cancerous cells, with the goal of identifying genes that were upregulated or downregulated in the cancerous cells as compared to the non-cancerous cells. They showed that VSN allowed more upregulated and downregulated genes to be identified than other transformations without loss of specificity.

2.3.4 Statistical tests for identifying peptides with significantly different signal intensities in different samples

The steps described in Section 2.3.3—normalization and the elimination of variance-versus-mean dependence—ensure that microarray data are suitable for identifying biological patterns. This section, along with Sections 2.3.5 and 2.3.6, describe methods for actually identifying those patterns.

The most basic biological question that can be answered using microarray data is, "which spots have statistically significantly different intensities in the different samples?" The exact interpretation of different

spot intensities depends, of course, on the type of microarray. For DNA microarrays, this means finding genes that are differentially expressed between two samples; for kinome microarrays, this means finding peptides that are differentially phosphorylated. In this section, statistical methods for answering this question are described.

Statistical tests can be divided into two categories based on assumptions about the distribution of the data being tested. Parametric tests assume that the data come from a particular probability distribution (such as a normal distribution). Nonparametric tests make no such assumption; however, their statistical power is usually less than parametric tests. As the VSN transformation results in an approximately normal distribution, the remainder of this section focuses on parametric significance tests.

The statistic used for determining whether a particular peptide is differentially phosphorylated depends on how many samples are being compared [Cui and Churchill, 2003]. It may occasionally be of interest to determine whether the level of phosphorylation of a given peptide is the same for all the samples in the experiment. In this case, analysis of variance (ANOVA) [Stekel, 2003] may be used; for a given peptide, ANOVA will indicate whether the means of the technical replicates for each sample are not all equal. More often, however, it is of interest to compare pairs of samples. For instance, suppose that an experiment is performed in which tissue samples are taken both from patients afflicted with a certain type of cancer, and from healthy controls. These samples are then analyzed on a kinome microarray, and the experimenter wishes to determine whether a given peptide has a different level of phosphorylation in the cancer patients versus the healthy controls. Statistically, the null hypothesis H_0 states that there is no difference in the phosphorylation of this peptide between the two groups—that is, that the means of the two groups are equal. Conversely, the alternative hypothesis H_A states that there is a difference (unequal means). A standard t-test can be used to determine whether the null hypothesis should be rejected in favour of the alternative hypothesis for a given peptide. The two-sample t-statistic can be calculated using the formula $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$, where n_1 and n_2 are the sample sizes, $\bar{x}_1$ and $\bar{x}_2$ are the sample means, and s_1^2 and s_2^2 are their variances [Stekel, 2003]. A P-value can then be obtained by comparing the value of t with a table for the t-distribution. This P-value can be compared to some preselected significance value α, often 0.05 or 0.01, to determine whether or not the null hypothesis should be rejected.

The t-test can also be used in experiments where there is not just a single treatment group and a single control group. For example, in a time course experiment, it may be of interest to use a t-test to compare each sample to the one taken at time 0, or to the sample taken at the previous time point. It should also be noted that t-tests can be used to compare two samples (or groups of samples) for which neither could appropriately be described as a "control". For instance, suppose that a hypothetical study involves comparing the effects of two novel drugs on signaling pathways in rats. While both drugs might separately be compared to a placebo (or no treatment at all), they also might be compared with one another. In this comparison, a t-test could be used to compare them even though neither is a control.

Since all arrays used in a given kinome microarray experiment have the same layout of peptides, a given

technical replicate on an array has a corresponding "partner" on another array by virtue of them occupying the same position. For instance, the red spot in the top-right corner of Figure 2.7 would have a corresponding spot on another array. Because of this property, a paired t-test can be used instead of an unpaired test. In this case, rather than calculating the means $\bar{x}_1$ and $\bar{x}_2$, one sample from a given pair is subtracted from the other sample. The mean of these differences is then calculated to produce a single average $\bar{x}$. The t statistic is then calculated using the formula $t = \bar{x}/(s/\sqrt{n})$, where s is the standard deviation of the aforementioned differences and n is the number of pairs.

The t-test indicates only whether there is a difference in mean phosphorylation intensity between two samples; it does not say anything about the degree of difference. For this, a fold-change (FC) ratio is often calculated. The appropriate formula for calculating the FC ratio differs depending on the normalization method used. In the case of untransformed data, it can simply be calculated as $\bar{x}_1/\bar{x}_2$. As described in Chapter 7, it is appropriate to use $2^{\bar{x}_1-\bar{x}_2}$ for data that has been transformed using VSN.

2.3.5 Identifying differentially modulated signaling pathways

As described in Section 2.2.4, the most basic question that can be answered using kinome microarrays is, "Which peptides are differentially phosphorylated in the treatment condition relative to the control condition?" However, this information is of little use if not subjected to further interpretation. As discussed in Section 2.1.6, individual proteins in the cell are components of signaling pathways, and it is the modulation of these pathways that ultimately affects the physiology of the cell. Therefore, the larger question when analyzing kinome microarray data is, "What biological pathways are upregulated or downregulated in the treatment condition compared to the control condition?"

For example, suppose that a hypothetical kinome array contains peptides corresponding to the proteins involved in the carbohydrate metabolism pathway discussed in Section 2.1.6. If this array were exposed to cell lysate, these proteins would be expected to exhibit very different levels of phosphorylation if the cells were exposed to insulin prior to lysis compared to if they were not. Specifically, in the presence of insulin, one would expect reduced phosphorylation of both the peptide representing glycogen synthase and the peptide representing glycogen synthase kinase, but increased phosphorylation of the peptide representing AKT.

While t-tests can be performed to determine whether individual peptides are differentially phosphorylated to a statistically significant degree (see also Section 2.3.4) in a treatment versus a control, determining whether an entire pathway is differentially modulated requires comparing the relative phosphorylation status of each peptide in that pathway. This can be done using a tool such as InnateDB [Lynn et al., 2008, Breuer et al., 2013]. As its name suggests, InnateDB is primarily a database of genes, proteins, and pathways associated with the innate immune system. In addition to containing its own manually-curated data, InnateDB also includes data from other pathway databases, including KEGG [Kanehisa and Goto, 2000, Kanehisa et al., 2006, 2010], the National Cancer Institute-Nature Pathway Interaction Database [Schaefer et al., 2009], the Integrating Network Objects with Hierarchies Pathway Database [Yamamoto et al., 2011],

NetPath [Kandasamy et al., 2010], and Reactome [Joshi-Tope et al., 2005, Croft et al., 2011]. In addition to storing data, InnateDB also contains several analysis tools. One of these tools, called "Pathway analysis", allows the user to upload a table containing a list of genes or proteins, along with a quantitative measure of the upregulation or downregulation of each. While designed for use with gene expression data from DNA microarrays, InnateDB is equally applicable to data from kinome arrays. Once the table has been uploaded, InnateDB returns a list of biological pathways that are significantly upregulated or downregulated, along with a P-value for each. From this information, a broader picture of changes in cellular physiology in the treatment condition relative to the control condition can be ascertained.

2.3.6 Clustering

Another important question when performing kinome microarray experiments is, "How similar are the various samples to one another in terms of their kinome profiles?" Clustering techniques, which can be used to answer this question, are described in this section. In particular, the use of distance metrics to measure the similarity of the phosphorylation profiles for a given pair of samples is described, along with two specific clustering techniques: hierarchical clustering and principal component analysis.

Distance metrics

In the context of kinome microarrays, clustering techniques typically require some way to numerically measure the similarity between the kinome profiles of a given pair of samples. In a kinome microarray experiment, the phosphorylation intensity data for a given sample can be represented as a vector $x = (x_1, x_2, \ldots, x_n)$, where x_i represents the average normalized intensity value for peptide i. The similarity between two samples x and y (for simplicity, the symbols x and y are overloaded to represent both the samples themselves and the vectors representing the data from those samples) can be evaluated using a distance metric $d(x, y)$. The smaller the value of $d(x, y)$, the more similar the kinome profiles of x and y.

In order to be a distance metric, the function d must satisfy the following properties:

1. $d(x, y) = d(y, x)$ for all x and y (symmetry);
2. $d(x, y) \geq 0$;
3. $d(x, x) = 0$; and
4. $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality).

There are several possible distance metrics. One is Euclidean distance, which represents the distance between two points in n-dimensional space (recall that n is the number of peptides). The formula for calculating Euclidean distance is very simple: $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ [Stekel, 2003]. One problem with Euclidean distance is that it is not scale invariant, meaning that one sample that exactly mirrors the behaviour of another, except on a different scale, would have a large Euclidean distance. For instance, suppose that

$n = 4$, $x = (1, 2, 3, 4)$, and $y = (4, 8, 12, 16)$. Although perfectly correlated, x and y have different scales and thus would have a large Euclidean distance. Thus, it is critical to first normalize the data (see Section 2.3.3) before using Euclidean distance as a distance metric.

A second distance metric is $\arccos(1 - r)$, where r is the standard (Pearson) correlation coefficient between x and y. Another function involving r, $\sqrt{1 - r}$, is also a distance metric. However, a similar function, $1 - r$, does not satisfy the triangle inequality and thus is not a metric. $1 - r$ can still be used as a distance, but it does not satisfy all four requirements of a distance metric as defined above.

Hierarchical clustering

Hierarchical clustering is a commonly-used clustering method, probably because it is both conceptually simple and computationally inexpensive. It is performed using the following iterative algorithm [Eisen et al., 1998]. First, choose a distance metric d. Second, calculate $d(x, y)$ for all pairs of samples x and y. Third, find the pair of samples having the smallest distance. These samples are merged to form a "combined sample" $A = \{x, y\}$. At this point, x and y are deleted from the list of samples, and A is added. The distance between A and each of the remaining samples is then computed, and the above procedure is again followed: the two samples with the smallest distance between them are combined. One or both of these samples may be "combined samples"; for instance, if the pair of samples with the smallest distance includes $A = \{x, y\}$ and $B = \{i, j, k\}$, then the result would be $C = \{x, y, i, j, k\}$.

Although the distance metric defines the distance between two individual samples, it does not indicate how to determine the distance between two combined samples. For this, a linkage method must be chosen. An example of a linkage method is average linkage, which is calculated as $d(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$. Clustering implementations that use average linkage as the linkage method sometimes call the overall procedure the unweighted pair group method with arithmetic mean (UPGMA). Another example is complete linkage, in which the distance between two combined samples is equal to the largest distance between an element from the first combined sample and an element from the second: $d(A, B) = \max_{a \in A, b \in B} d(a, b)$. In the McQuitty linkage method, if clusters A and B are being joined to create a new cluster C, then the distance between C and some other cluster D is computed as $d(C, D) = (d(A, D) + d(B, D))/2$ [McQuitty, 1966].

The end result of the hierarchical clustering algorithm can be represented as a dendrogram, a tree-like diagram in which samples are represented as leaves, and samples on the same branch are more similar than samples on different branches. Examples of dendrograms created using hierarchical clustering can be found in Figures 9.1A, 9.2A, 9.3A and 10.3.

Despite its benefits, hierarchical clustering also has some deficiencies. Because each iteration of the algorithm involves finding the pair of samples with the smallest distance, the decisions made may be optimal at the local level, but not at the global level. Hierarchical clustering is thus classified as a greedy algorithm. Critically, poor decisions made at the beginning of the algorithm (for example, joining together two samples

whose phosphorylation profiles appear similar, but really are not) can compromise the accuracy of the algorithm throughout, potentially leading to an inaccurate final result. A second drawback is that the results can be difficult to interpret. This is a consequence of the iterative clustering process, in which there is first a cluster containing two samples, and then a cluster containing three samples, and so on, finally resulting in a cluster containing all of the samples in the experiment. The presence of so many clusters may make it difficult to evaluate which ones have biological meaning. Besides those mentioned above, there are other, more minor, problems with hierarchical clustering, an empirical investigation of which is given by Morgan and Ray [1995].

Principal Component Analysis

The dimensionality of kinome microarray data is high—each sample has n values associated with it, where n, the number of unique peptides on the array, is on the order of hundreds or thousands. This level of dimensionality can make the data difficult to visualize and analyze. Fortunately, often there are sets of variables (peptides) for which the variables in that set are correlated with one another, and can therefore be considered as if they were just a single variable. Principal component analysis (PCA) is a method that uses matrix operations to reduce the dimensionality of a dataset, distilling the original set of variables into a smaller set of "principal components", each of which captures a portion of the variability in the data. The amount of variability captured by a given principal component depends on the relationships among the original variables. Be definition, the first principal component captures the most variability, followed by the second principal component, and so on. Let m represent the number of principal components for which the proportion of explained variation is high (say, greater than 10%). If $m \ll n$, then this represents a substantial reduction in dimensionality compared to the original set of data. However, if $m > 3$, then the data can still be difficult to visualize. In such cases, a common practice is to perform visualization using only the first two or three principal components, in which case visualization would be in the form of a two-dimensional or three-dimensional scatterplot, respectively. This approach is particularly useful in cases where the first two or three principal components capture a large portion of the variability in the original data, rendering the remaining principal components less important. If two samples are close together in these scatterplots, then those samples have similar kinome profiles. Examples of the three-dimensional visualization of kinome microarray data after performing PCA can be found in Figures 9.2C and 9.3C.

Chapter 3

Computational prediction of eukaryotic phosphorylation sites

Brett Trost and Anthony Kusalik

This is the first of four papers that relate to the design of kinome microarrays. As mentioned in Chapter 1, there are many organisms for which few phosphorylation sites have been experimentally identified, making it difficult to design kinome arrays suitable for studying them. As such, computational methods are required in order to predict potential sites. In this paper, the challenges involved with the computational prediction of phosphorylation sites are reviewed, along with potential solutions to those problems. Forty existing prediction methods are compared in terms of several characteristics, including the machine-learning method used, the number of residues surrounding the phosphorylation sites that are used in the model, whether or not structural information is used, whether the models are kinase-specific or designed for kinases in general, and the sources of training data used. The paper attempts to be useful both to developers of predictors (who might wish to improve upon previous prediction methods) and to biologists (who may face the problem of choosing an appropriate predictor for their specific biological application). Several future directions in the field of phosphorylation site prediction are also discussed.

Citation

B. Trost and A. Kusalik. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* 27(21):2927–2935, 2011.

Author contributions

Brett Trost performed the research and wrote the paper. Anthony Kusalik supervised the research and helped edit and revise the paper.

3.1 Abstract

Motivation: Kinase-mediated phosphorylation is the central mechanism of post-translational modification to regulate cellular responses and phenotypes. Signaling defects associated with protein phosphorylation are linked to many diseases, particularly cancer. Characterizing protein kinases and their substrates enhances our ability to understand and treat such diseases and broadens our knowledge of signaling networks in general.

While most or all protein kinases have been identified in well-studied eukaryotes, the sites that they phosphorylate have been only partially elucidated. Experimental methods for identifying phosphorylation sites are resource-intensive, so the ability to computationally predict potential sites has considerable value.

Results: Many computational techniques for phosphorylation site prediction have been proposed, most of which are available on the web. These techniques differ in several ways, including the machine learning technique used; the amount of sequence information used; whether or not structural information is used in addition to sequence information; whether predictions are made for specific kinases or for kinases in general; and sources of training and testing data.

This review summarizes, categorizes, and compares the available methods for phosphorylation site prediction, and provides an overview of the challenges that are faced when designing predictors and how they have been addressed. It should therefore be useful both for those wishing to choose a phosphorylation site predictor for their particular biological application, and for those attempting to improve upon established techniques in the future.

3.2 Introduction

Phosphorylation is the most widespread post-translational modification in eukaryotes and plays a crucial role in the regulation of virtually every cellular behavior, including DNA repair [Wood et al., 2009], environmental stress response [Wang et al., 2010], regulation of transcription [Uddin et al., 2003], apoptosis [Zhang and Johnson, 2000], cellular motility [Ressurreição et al., 2011], immune response [Kim and Lee, 2011], metabolism [Bu et al., 2010], and cellular differentiation [Lian et al., 2010]. Historically, novel phosphorylation sites have been discovered primarily through the use of low-throughput biological techniques.

With the advent of site-directed mutagenesis, for instance, many labs started using this technique to characterize specific phosphorylation events [e.g., Meier et al., 1997]. Unfortunately, such techniques are time-consuming, tedious, and expensive to perform. More recently, a high-throughput technique—mass spectrometry—has greatly accelerated the identification of novel sites. For example, Huttlin et al. [2010] used mass spectrometry to map the phosphoproteome of nine different mouse tissues, identifying nearly 36,000 distinct phosphorylation sites. While useful, this technique has important limitations: it cannot identify the

protein kinase(s) responsible for catalyzing the phosphorylation of a given site; many phosphorylation sites are modified at substoichiometric levels, with the unphosphorylated form sometimes preventing detection of the phosphorylated form; many interesting proteins are present at very low levels, making them difficult to detect through mass spectrometry; breaking open cells can place kinases together with substrates they would not normally encounter, potentially resulting in the detection of phosphorylation events that would not occur *in vivo*; technical challenges exist that sometimes make pinpointing exact phosphorylation sites difficult [Boersema et al., 2009]; and perhaps most importantly, mass spectrometry requires very expensive instruments and specialized expertise that are not available in typical laboratories.

Given the limitations associated with both low-throughput and high-throughput biological techniques for identifying novel phosphorylation sites, computational approaches have become increasingly popular. Such techniques require a protein sequence as input, and output some numerical measure of the likelihood that each serine, threonine, or tyrosine (S/T/Y) residue in that sequence is a phosphorylation site. For example, Slaugenhaupt et al. [2001] found that the mutation R696P in the protein encoded by the *IKBKAP* gene causes familial dysautonomia, with the hypothesized mechanism being disruption of the phosphorylation of T699—a site predicted to be phosphorylated by NetPhos [Blom et al., 1999], the first phosphorylation site prediction tool. Prediction programs are often used to narrow down the list of possible phosphorylation sites in a protein of interest, with the predictions subsequently verified using biological experiments. For instance, Fan et al. [2009] used a combination of several predictors to identify seven putative sites phosphorylated by protein kinase C in the transient receptor potential vanilloid 4 (TRPV4) ion channel. When three of these sites were separately mutated to alanine, the resultant proteins exhibited markedly reduced activation in response to protein kinase C compared to wild-type TRPV4.

To the authors' knowledge, four review articles have previously been published that included significant discussion of computational phosphorylation site prediction. Kobe et al. [2005] provided a brief review of this field, along with a detailed discussion of the structural bases of protein kinase specificity. Hjerrild and Gammeltoft [2006] reviewed both computational and biological aspects of phosphoproteomics, while Miller and Blom [2009] briefly summarized some of the literature on phosphorylation site prediction and provided a protocol for the use of their NetPhos [Blom et al., 1999, Hjerrild et al., 2004, Blom et al., 2004] family of tools. Most recently, Xue et al. [2010] reviewed phosphorylation site databases, prediction tools, and miscellaneous software associated with phosphorylation sites, and also compared the performance of a subset of available predictors.

Compared to the above reviews, here we concentrate more specifically on the methodologies employed by the various prediction tools, taking a comparative approach to examining the issues and challenges associated with computational phosphorylation site prediction. Section 3.3 of this review provides a brief overview of the available methods, while Section 3.4 compares and discusses the tools with respect to different aspects of their methodologies. Section 3.5 comments on some of the challenges that remain in the field, and Section 3.6 gives some concluding remarks.

3.3 An overview of current tools for phosphorylation site prediction

If significant updates to existing methods are treated separately, then there have been nearly 40 methods for the computational prediction of phosphorylation sites described since 1999. This total excludes tools that predict sites for more than one type of post-translational modification [e.g., Schwartz et al., 2009, Basu and Plewczynski, 2010] and methods based on simple motifs [discussed by Xue et al., 2010]. Unlike Xue and co-authors, however, we include techniques for which no web implementation is available, as they can be valuable sources of ideas for developers of future tools.

A list of currently available phosphorylation site prediction tools is given in Table 3.1. Each tool is categorized with respect to several important attributes, including the machine learning technique(s) used (described further in Section 3.4.1); the number of residues surrounding the phosphorylation site that are taken into account (Section 3.4.2); whether the method uses only sequence information or also uses structural information (Section 3.4.3); whether the tool includes models specific to particular kinases or kinase families (Section 3.4.4); and the source(s) of known phosphorylation sites used for training and testing (Section 3.4.5). We avoid comparing the performance of each tool, for several reasons: some of the tools discussed do not have web implementations or have websites that are no longer accessible; different tools use different performance measures or were tested on different datasets; it would not be meaningful to compare kinase-specific with non-kinase-specific tools; and performance comparisons have been done elsewhere for some tools [Xue et al., 2010]. However, we do discuss two important issues regarding predictive performance—the creation of standardized testing datasets (Section 3.5.1), and balancing sensitivity and specificity (Section 3.5.3).

3.4 Comparing and contrasting the available tools

3.4.1 Machine learning methods

To provide tight control of cellular processes, a protein kinase catalyzes the phosphorylation of a given S/T/Y residue only if the amino acids around that residue fit a specific, yet flexible, pattern [Diks et al., 2007]. Sequence motifs that describe these patterns, such as those in the PROSITE database [Sigrist et al., 2002], are neither sensitive [see Blom et al., 1999] nor specific (the PROSITE motif for the protein kinase C recognition sequence, for instance, is [ST]-x-[RK], which would be expected to occur often at random). The poor specificity and sensitivity of motifs means that accurate prediction of phosphorylation sites requires the use of machine learning methods, which can identify more complex and subtle patterns. As Table 3.1 shows, many different machine learning methods have been used, including artificial neural networks (ANN), decision trees, genetic algorithms, position-specific scoring matrices (PSSM), and support vector machines (SVM).

Table 3.1: Currently available phosphorylation site prediction tools. Column headings are as follows: name, the name of the tool; technique, the machine learning technique used; residues, the number of residues flanking (and including) the phosphorylated residue that are used in the tool's predictions; 1D/3D, whether only sequence information is used (1D) or structural information is used as well (3D); K-spec, yes if the tool makes predictions for specific kinases or kinase families, and no otherwise; data, the source of the known phosphorylation sites used for training and testing; reference, the paper describing that tool; website, the address of that tool's web implementation (if applicable). Due to space considerations, only one reference per tool is included in the table; tools described by more than one paper include ScanSite (also described in Obenauer et al. [2003]), NetPhosK [Blom et al., 2004], PredPhospho [Ryu et al., 2009], GPS 1.0 [Xue et al., 2005], KinasePhos 1.0 [Huang et al., 2005a], PHOSIDA [Gnad et al., 2011], PhosPhAt [Durek et al., 2010], Predikin 2.0 [Saunders and Kobe, 2008], and Musite [Gao et al., 2010]. Abbreviations: ANN, artificial neural network; BP, Bayesian probability; CRF, conditional random fields; DT, decision tree; GA, genetic algorithm; IHE, in-house experiments; LIT, literature; LR, logistic regression; MC, Markov clustering; MP, meta-predictor; PB, PhosphoBase; P.ELM, Phospho.ELM; PPA, PhosPhAt database; PS, PhosphoSitePlus; PSSM, position-specific scoring matrix; SA, structural analysis; SP, Swiss-Prot (or UniProt); STAT, statistical method; SVM, support vector machine; TAIR, The *Arabidopsis* Information Resource database. *No name was given to these tools by their authors. †Exact range of lengths not explicitly stated. §Varies depending on individual predictors used (see text).

Name	Technique	Residues	1D/3D	K-spec?	Data	Reference	Website
NetPhos	ANN	9–33	3D	No	PB	Blom et al., 1999	cbs.dtu.dk/services/NetPhos
Scansite	PSSM	15	1D	Yes	PB	Yaffe et al., 2001	scansite.mit.edu
Predikin 1.0	SA	7	3D	Yes	PB	Brinkworth et al., 2003	predikin.biosci.uq.edu.au
rBBFNN	ANN, DT	9–11	1D	No	PB	Berry et al., 2004	(no web implementation available)
DISPHOS	LR	25	3D	No	PB, SP	Iakoucheva et al., 2004	www.dabi.temple.edu/disphos
NetPhosK	ANN	9–33	3D	Yes	Many	Hjerrild et al., 2004	cbs.dtu.dk/services/NetPhos
PredPhospho	SVM	†	1D	Yes	PB, SP	Kim et al., 2004	(website no longer accessible)
PHOSITE	PSSM	†	1D	Yes	PB	Koenig and Grabe, 2004	(website no longer accessible)
GPS 1.0	PSSM, MC	7	1D	Yes	P.ELM	Zhou et al., 2004	gps.biocuckoo.org
*	Many	9	1D	No	PB	Senawongse et al., 2005	(no web implementation available)
KinasePhos 1.0	HMM	9	1D	Yes	PB, SP	Huang et al., 2005b	kinasephos.mbc.nctu.edu.tw
*	SVM	9	3D	Yes	SP	Plewczyński et al., 2005	(no web implementation available)
PPSP	BP	9	1D	Yes	P.ELM	Xue et al., 2006	ppsp.biocuckoo.org
pkaPS	SA	42	3D	Yes	P.ELM	Neuberger et al., 2007	mendel.imp.ac.at/sat/pkaPS
*	STAT	2–4	1D	Yes	LIT, P.ELM	Moses et al., 2007	(no web implementation available)
NetPhosYeast	ANN	†	1D	No	LIT, SP	Ingrell et al., 2007	cbs.dtu.dk/services/NetPhosYeast
NetworKIN	ANN, PSSM	9–33	3D	Yes	P.ELM	Linding et al., 2007	networkin.info
KinasePhos 2.0	SVM	9	1D	Yes	P.ELM	Wong et al., 2007	kinasephos2.mbc.nctu.edu.tw
GANNPhos	Many	25	1D	No	P.ELM, SP	Tang et al., 2007	(no web implementation available)
PHOSIDA	SVM	13	1D	No	PHOSIDA	Gnad et al., 2007	phosida.de
PhosPhAt	SVM	13	1D	No	PPA	Heazlewood et al., 2008	phosphat.mpimp-golm.mpg.de
IEPP	BP	31	1D	Yes	P.ELM	Wang et al., 2008a	(no web implementation available)
AutoMotif	SVM	9	1D	Yes	SP	Plewczyński et al., 2008	(website no longer accessible)
PhoScan	PSSM	25	1D	Yes	P.ELM	Li et al., 2008b	bioinfo.au.tsinghua.edu.cn/phoscan
MetaPredPS	MP	§	§	Yes	Many	Wan et al., 2008	metapred.biolead.org/MetaPredPS
SiteSeek	Many	†	1D	Yes	Many	Yoo et al., 2008	(no web implementation available)
SMALI	PSSM	7	1D	Yes	IHE	Li et al., 2008a	lilab.uwo.ca/SMALI.htm
Predikin 2.0	HMM, SA	7	3D	Yes	P.ELM, SP	Saunders et al., 2008	predikin.biosci.uq.edu.au
GPS 2.0	PSSM, GA	15	1D	Yes	P.ELM	Xue et al., 2008	gps.biocuckoo.org
CRPhos	CRF	9	1D	Yes	P.ELM	Dang et al., 2008	www.ptools.ua.ac.be/CRPhos
Phos3D	SVM	13	3D	Yes	P.ELM	Durek et al., 2009	phos3d.mpimp-golm.mpg.de
*	SVM	25	1D	No	PPA, TAIR	Gao et al., 2009a	(no web implementation available)
PostMod	PSSM	7–101	1D	Yes	P.ELM	Jung et al., 2010	pbil.kaist.ac.kr/PostMod
PPRED	PSSM, SVM	7–15	1D	No	P.ELM	Biswas et al., 2010	ashiskb.info/research/ppred
*	SVM	9–15	3D	No	P.ELM	Swaminathan et al., 2010	(no web implementation available)
BAE	STAT	11	1D	Yes	P.ELM	Yu et al., 2010	(no web implementation available)
PAAS	PSSM	†	1D	Yes	P.ELM	Sobolev et al., 2010	(website no longer accessible)
*	STAT, SVM	9	1D	Yes	P.ELM	Li et al., 2010	cmbi.bjmu.edu.cn/huphospho
Musite	SVM	†	1D	Yes	Many	Gao and Xu, 2010	musite.sourceforge.net
GPS 2.1	PSSM, GA	3–31	1D	Yes	P.ELM	Xue et al., 2011	gps.biocuckoo.org

Perhaps the simplest machine learning technique is the PSSM, which is a matrix in which rows represent amino acids and columns represent positions in a multiple sequence alignment. In the simplest possible PSSM, a given matrix element would contain the frequency of a given amino acid in a given position, although more complex variations are usually developed in practice [e.g., Koenig and Grabe, 2004, Li et al., 2008b]. PSSMs are easy to understand and construct, but are unable to detect patterns in which combinations of amino acids are important [Blom et al., 1999]. PSSMs can, say, express the idea that proline promotes phosphorylation when found at position +1 (where position 0 is the phosphorylation site) and arginine promotes phosphorylation when found at −2, but cannot express the idea that both occurring at the same time prevents phosphorylation.

In contrast, machine learning techniques like ANNs and SVMs—two of the most popular methods used by phosphorylation site prediction tools—can capture more complex patterns [Blom et al., 1999]. This comes at the cost of added complexity. ANNs, in particular, are often regarded as "black boxes" in which the classification function is essentially inscrutable. Some methods strike a balance between the simplicity of PSSMs and the opaqueness of ANNs. Xue et al. [2006], for example, proposed a method based on Bayesian probability that is more expressive than PSSMs, but is more easily interpreted biologically and mathematically than ANNs.

An interesting point of discussion is, what do the machine learning methods actually model? In other words, do they model the actual biological mechanisms underlying protein kinase recognition, or do they merely recognize patterns? For the majority of methods listed in Table 3.1, and certainly for those that consider only sequence information (see also Section 3.4.3), we would argue that it is the latter. This is not meant to denigrate these methods: clearly, recognizing patterns in sequence information is useful, both in the field of phosphorylation site prediction and elsewhere. Most tools utilizing structural information fall somewhere between the two categories mentioned above. For example, although DISPHOS [Iakoucheva et al., 2004] uses secondary structure predictions as features, it would be better described as recognizing patterns than as modeling biological mechanisms. On the other end of the scale, pkaPS [Neuberger et al., 2007] extensively models the kinase-substrate interaction. While pattern recognition has resulted in much success, it is plausible that more closely modeling the underlying biology of substrate recognition will result in the greatest gains in predictive performance.

3.4.2 Amount of sequence information used

Phosphorylation site prediction tools vary widely in the number of residues surrounding the phosphorylation site that are taken into account. At one extreme, PostMod [Jung et al., 2010] was designed to consider up to 101 residues (between positions −50 and +50), whereas Predikin 1.0 [Brinkworth et al., 2003] considers just seven. The number of residues considered is important because too few means information useful for making predictions gets ignored, while too many will decrease the signal-to-noise ratio. Using many residues can also make some machine learning methods computationally intractable [Biswas et al., 2010].

Several strategies have been used to determine the optimum number of residues. First, it has been argued that the optimum should be consistent with the number of residues in physical contact with the kinase [Blom et al., 1999]. An early report stated that 9–12 residues surrounding the phosphorylation site are likely to physically contact the kinase [Songyang et al., 1994], an estimate consistent with the number of residues used by many prediction methods. Depending on the three-dimensional structure of the substrate, however, the 9–12 residues contacted by the kinase may not be the same as the 9–12 residues surrounding the phosphorylation site in the linear sequence. Residues not in contact with the kinase may also affect its binding by influencing the charge or hydrophobicity of the microenvironment, or by affecting the conformation of residues in contact with the kinase. As such, the number of residues that physically contact the kinase may not reliably indicate the number of residues that should be used for making predictions.

Second, some authors have empirically tested various numbers of residues, and then chosen the number that gives the best predictive performance. The authors of PostMod [Jung et al., 2010] tried between seven and 101 residues, and found that 41 resulted in the best accuracy. Other authors reported much smaller optima, with Blom et al. [1999] suggesting between nine and 11 and Biswas et al. [2010] reporting 15. Given that reported optima are inconsistent, researchers should use caution when applying previously reported empirical optima for developing future methods.

Third, Neuberger et al. [2007] examined how residues around phosphorylation sites compare with residues in general proteins with respect to two properties—hydrophobicity and flexibility. Figure 1 of their paper plots position (40 residues upstream and downstream of the phosphorylation site) versus deviation from baseline values, and shows that each property deviates substantially from baseline values near phosphorylation sites, and then gradually returns as one gets farther from a given site. The authors found that both properties deviate significantly between positions -18 and $+23$, and thus advocate the use of these 42 residues. Because this experiment was done only for protein kinase A, it is not known whether its results generalize to other protein kinases.

The lack of agreement among the three strategies described above may be due to differences among kinases (some kinases may use more residues as a recognition sequence than others) and/or among machine learning methods (some methods may handle greater dimensionality—in other words, longer sequences—better than others). The lack of agreement could also be related to effect size. For example, a residue at position -20 may have a real, but very small, effect on phosphorylation—an effect that might be ignored by some authors, but not others. As the most appropriate number of residues remains unclear, a rigorous investigation of this issue would be invaluable for developers of future tools, especially if consideration was given to the particular machine learning technique used and the particular kinase under consideration.

3.4.3 Use and non-use of structural information

The structural basis of protein kinase-catalyzed phosphorylation has been examined in numerous studies. For example, Dunker et al. [2002] reported that phosphorylation sites are frequently found in disordered

regions (a fact exploited by the authors of the NETPHOS prediction tool [Iakoucheva et al., 2004]), while Kitchen et al. [2008] described the degree to which electrostatic interactions stabilize phosphorylated residues. Further, a review by Kobe et al. [2005] examined the structural determinants of protein kinase specificity.

The degree to which the substrate's three-dimensional structure affects kinase specificity is unclear. Studies that find correlations between these two variables, like those cited above, suggest that the substrate's structure is important in the recognition process. Conversely, short peptides containing known phosphorylation sites can be recognized with similar kinase-catalyzed kinetics as the corresponding intact protein (Zetterqvist et al. [1976], Kemp et al. [1977]; see also Houseman et al. [2002] and Löwenberg et al. [2005]), suggesting a minor role for structure. Despite this, it seems plausible that the use of structural information can play at least some role in increasing the accuracy of phosphorylation site prediction tools. Although lack of structural data remains an obstacle, the amount of structural information about phosphorylation sites is growing rapidly. The most recent version of the Phospho3D database [Zanzoni et al., 2011], for instance, contains structural information for over 1700 sites, nearly 11 times the number contained in the previous version [Zanzoni et al., 2007].

Table 3.1 shows that approximately one quarter of tools utilize information regarding the three-dimensional structure of the kinase and/or its substrate, whereas the other tools use only primary sequence information. Blom et al. [1999], in addition to devising a method based only on primary sequence, superimposed the structures of 12 different tyrosine phosphorylation sites, and found that nine of them had a common conformation, while the other three shared a second conformation. In contrast, non-phosphorylated tyrosine residues exhibited a wide range of conformations. They also determined that phosphorylated residues were generally more flexible than average, consistent with the hypothesis that high flexibility would be required to fit into a kinase's active site. While conformation and flexibility thus seemed like two structural features that could increase prediction accuracy, the authors' sequence-based method outperformed their structure-based method, although the latter did make more accurate predictions for a few atypical tyrosine phosphorylation sites. In contrast, Durek et al. [2009] found that, for several different kinase families, adding structural information to a sequence-only model resulted in a modest but consistent increase in predictive performance, showing that the use of structural information can add discriminatory power.

Given that sequence-only methods, by definition, ignore information about the kinase-substrate interaction, the upper limit to their accuracy is likely less than the upper limit of structure-based methods. As structural information becomes available for more and more phosphorylation sites, structure-based methods will continue to improve.

3.4.4 Kinase-specific versus non-kinase-specific tools

Most phosphorylation site prediction programs are kinase-specific, as they require as input both a protein sequence and the name of a protein kinase, and output some measure of the likelihood that each S/T/Y residue in the sequence is phosphorylated by the chosen kinase. In contrast, a few tools require only a

protein sequence as input, and output the likelihood that each S/T/Y residue is phosphorylated by any kinase. Kinase-specific tools can be further divided based on whether they make predictions for individual kinases (e.g., NetPhosK [Blom et al., 2004] and pkaPA [Neuberger et al., 2007]) or for kinase families (e.g., SiteSeek [Yoo et al., 2008] and PAAS [Sobolev et al., 2010]). Part of the motivation for making predictions for kinase families is that some individual kinases have very few target sites known, making the training component of machine learning difficult. As kinases from the same family will likely have similar recognition sequences [Kim et al., 2004], their known target sites can be combined, resulting in a model that utilizes much more information than if kinases from the same family were modeled separately.

How do the accuracies of non-kinase-specific tools compare with those of kinase-specific tools? Given the issues involved in comparing the performance of different tools (see Section 3.4.5), this question is more difficult to answer than it would appear. It has been claimed that, since there is no "average" phosphorylation site, only kinase-specific predictors should be able to achieve good accuracy [Neuberger et al., 2007]—an argument with considerable logical appeal. Indeed, most users will likely be interested in particular biological pathways (and thus particular kinases), making kinase-specific tools an ideal choice. For applications in which the specific kinase is not a concern, the user could still take advantage of the higher accuracy of kinase-specific tools by aggregating the results from many kinase-specific predictions to make a general list of phosphorylation sites in the protein(s) of interest. On the other hand, non-kinase-specific tools may be able to detect phosphorylation sites for which the associated kinase is unknown—an advantage that may be of interest to some users. Additionally, non-kinase-specific tools have reported respectable performance, with accuracy rates approaching 80% [Swaminathan et al., 2010].

3.4.5 Training and testing data

Both positive (actual phosphorylated residues) and negative (actual non-phosphorylated residues) data are required for training and testing a particular prediction tool. This section discusses sources of positive and negative data, as well as the issue of fair performance comparisons.

Positive data

Several sources of known phosphorylation sites have been used. Most early prediction tools used either PhosphoBase [Blom et al., 1998, Kreegipuu et al., 1999], a database solely containing known phosphorylation sites, or Swiss-Prot, for which the annotation of a given protein includes its known phosphorylation sites. Authors using Swiss-Prot [e.g., Iakoucheva et al., 2004, Plewczyński et al., 2005] generally discard sites described as "hypothetical", "predicted", or "by similarity", choosing instead only experimentally confirmed sites. In 2004, the information from PhosphoBase was integrated into a new database called Phospho.ELM [Diella et al., 2004, 2008, Dinkel et al., 2011]. Most tools developed after 2004 have used Phospho.ELM, although there are exceptions: other databases that have been used (some of which are specialized in nature) are PhosphoSitePlus [Hornbeck et al., 2004], The *Arabidopsis* Protein Phosphorylation Site Database (Phos-

PhAt) [Heazlewood et al., 2008, Durek et al., 2010], The *Arabidopsis* Information Resource (TAIR) [Swarbreck et al., 2008], and PHOSIDA [Gnad et al., 2007, 2011]. Finally, a few authors searched the literature for known phosphorylation sites [e.g., Hjerrild et al., 2004, Moses et al., 2007].

Negative data

An ever-present difficulty in the field of phosphorylation site prediction concerns negative training and testing data. While experiments can verify that a particular residue can be phosphorylated, it would be difficult to prove definitively that a particular residue is not phosphorylated under any conditions. Thus, while databases such as Phospho.ELM and PhosphoSitePlus contain thousands of known phosphorylation sites, they do not contain sites known not to be phosphorylated.

To circumvent this problem, most authors make the assumption that any S/T/Y residue that has not been shown to be phosphorylated is a negative. While some of these residues will likely turn out to be positives as more phosphorylation sites are discovered, the majority of these are probably actual negatives, making this a reasonable, if imperfect, approach. Some authors [e.g., Neuberger et al., 2007] have gone a step further, requiring that the residue not be found in any phosphorylation site database and that it be found in a protein for which there exists at least one residue known to be phosphorylated by the kinase of interest. The assumption here is that, if a protein has at least one residue that is known to be phosphorylated, then the phosphorylation of that protein has been studied in at least some detail, making it less likely that its other S/T/Y residues are undiscovered phosphorylation sites.

Another approach is to use, as negative data, S/T/Y residues that are buried in the core of a particular protein [Blom et al., 2004]. This strategy relies on the assumption that buried residues would not be physically accessible to any kinase, thus reducing the number of so-called negatives that later turn out to be positives. A disadvantage of this approach is that it requires knowledge of the protein's tertiary structure, and only a small portion of proteins currently have solved structures (although the use of structure prediction programs could partially compensate for this). More importantly, however, this method's underlying assumption may not be entirely valid. In a detailed analysis of experimentally-verified phosphorylation sites, Jiménez et al. [2007] found that while phosphorylation sites are more solvent-exposed than the average residue, close to 15% have little solvent accessibility. Moreover, a site can be buried in one structure of a given protein, but unburied in another [Zhou et al., 2006, Durek et al., 2009]. Despite these caveats, choosing solvent-inaccessible residues as negatives currently seems like the most reliable approach to obtaining negatives for training and testing.

Performing fair comparisons of performance

While new prediction tools can improve upon previous ones in various ways, developers must usually show that a new tool offers an improvement in predictive performance. To perform a fair comparison, both the new method and existing methods must be tested using the same data. When testing existing tools, typically one has access only to the already-trained versions that are available on the web [Dang et al., 2008], and

since data that were used to train a given tool should not be used to test it [Dang et al., 2008], it can be difficult to identify suitable testing data.

Some authors have simply ignored this problem, comparing their tools' performance numbers (sensitivity, specificity, etc.) directly with those given in the papers describing previous tools, even though the testing data used may have been different. While having some value, such comparisons are certainly less informative than they could be.

Positive data appropriate for comparing new and existing tools can be obtained by collecting known phosphorylation sites added to a database after the publication of all existing methods [Wan et al., 2008]. If access is available to known phosphorylation sites that have not yet been deposited in the databases, they could be used as well. Note that while new known sites are required for comparing performance with existing methods, older known sites can still be used for training a new method.

Given how negative data are obtained (see Section 3.4.5), obtaining negative data appropriate for testing seems harder than for positive data (and interestingly, has been given little or no attention in the literature). Suppose that, as many have done, developer A focuses exclusively on predicting phosphorylation in humans. Since the entire human proteome is known, he might use all S/T/Y sites not known to be phosphorylated as negative data for training his method. If developer B later wishes to compare his new method to that of A, there would be no negative data available that were not used to train A's method, making a fair comparison impossible.

While requiring coordination among those in the phosphorylation site database and prediction community, possible solutions to these problems do exist, some of which are suggested in Section 3.5.1.

3.4.6 Other differences among the available tools

While Table 3.1 categorizes the tools in terms of important properties for which they vary, these categories do not capture all of their differences, and there are several tools that deviate from the norm in a notable way. For instance, Musite [Gao and Xu, 2010] is unique in that it is an open-source platform that allows the creation of a customized predictor, with the user able to choose different training and testing data, features, stringency thresholds, and so on. Other examples of tools that differ from the norm are given below.

While most tools were trained using known phosphorylation sites, ScanSite's [Yaffe et al., 2001, Obenauer et al., 2003] authors created an oriented peptide library, then incubated it with a given protein kinase. Phosphorylated peptides were separated from those that were not phosphorylated, and the former sequenced to determine the abundance of each amino acid at each position. The ScanSite program uses this information to output the likelihood that a given S/T/Y residue in its input sequence can be phosphorylated by that protein kinase. Although known phosphorylation sites were not used for training, known sites from PhosphoBase [Blom et al., 1998, Kreegipuu et al., 1999] were used for testing. More recently, Li et al. [2008a] developed SMALI, a tool which is similar to ScanSite but claims to have better accuracy.

Most tools require protein sequences as input, and output a score indicating the likelihood that a given

S/T/Y residue is a phosphorylation site. In contrast, Predikin 1.0 [Brinkworth et al., 2003] takes the sequence of an uncharacterized protein kinase as input, and outputs a 7-mer predicted to be its optimal recognition sequence. Predikin 2.0 [Saunders et al., 2008, Saunders and Kobe, 2008] improved upon the original's ability to output optimal kinase recognition sequences and also added the conventional functionality of scoring potential phosphorylation sites.

Finally, MetaPredPS [Wan et al., 2008] is currently the only meta-predictor, which is a type of tool that combines the classifications from several individual predictors in the hope of achieving better accuracy. MetaPredPS uses a weighted voting strategy to combine predictions from GPS 1.0, KinasePhos 1.0, NetPhosK, PPSP, PredPhospho, and Scansite (see Table 3.1 for references). Meta-predictors have also been successfully applied to other bioinformatics-related classification problems, including subcellular localization prediction [Shen et al., 2007, Liu et al., 2007], major histocompatibility complex-binding prediction [Trost et al., 2007, Karpenko et al., 2008, Wang et al., 2008b], and protein structure prediction [Ginalski et al., 2003]. Given that many different strategies can be used to combine the output of individual predictors, and that there exist dozens of individual tools for phosphorylation site prediction, there is likely room for additional work on meta-predictors in this field.

3.5 Future directions

In some respects, the field of phosphorylation site prediction is mature. As Table 3.1 shows, many different machine learning methods have been utilized; widely varying amounts of information (in terms of number of residues surrounding the phosphorylation site) have been incorporated into predictive models; many methods have been proposed in both the structure-based and sequence-based categories; several tools exist for both kinase-specific and non-kinase-specific predictions; and many sources of training and testing data have been utilized. In other respects, however, the field is immature. Three challenges that remain (to rigorously determine the optimum number of residues surrounding the phosphorylation site, to develop improved structure-based methods, and to develop additional meta-predictors) were discussed earlier in this review. A number of others were described by Xue et al. [2010]. Four additional challenges, which we feel are of particular significance, are described below.

3.5.1 Creating standardized testing datasets

Perhaps the most important challenge involves the development of standardized testing datasets. As described in Section 3.4.5, it is currently extremely difficult to properly compare the accuracy of different prediction tools—either by reading the papers describing them, or by testing them anew. Given the large number of phosphorylation site prediction programs already available, it is critical that authors of newly-developed tools be able to show a clear improvement in performance compared to older ones. A dataset containing both positive and negative data, half of which is designated for training (only) and half of which is designated for

testing (only), would be an invaluable resource, as it would provide a fair, standardized benchmark by which each tool could be judged. Such a database could be created if a laboratory were to use mass spectrometry to identify as exhaustively as possible the phosphorylation sites in an organism for which few sites are currently known. The "buried residue method" (see Section 3.4.5) could then be used to identify negatives. Unfortunately, this solution has an important limitation: mass spectrometry does not give information about the kinase that phosphorylates each site—information required by tools making kinase-specific predictions. Another solution would be for curators of phosphorylation site databases to designate a portion of all future data collected (from low-throughput or high-throughput sources) as "testing data", and for the developers of future tools to voluntarily refrain from using these data for training. This strategy could substantially improve the ability to compare phosphorylation site prediction methods.

3.5.2 Developing tools for a wider variety of organisms

Both the quantity of protein kinases and the types represented (tyrosine kinases, calmodulin-dependent kinases, etc.) differ substantially in different eukaryotes [Diks et al., 2007]. For instance, *Arabidopsis* encodes around twice as many protein kinases as does human [Manning et al., 2002, Champion et al., 2004], but does not encode any classical tyrosine kinases. In addition, the lower eukaryote *Plasmodium falciparum* encodes only a few dozen protein kinases, but some of these are of a type observed in few other organisms [Ward et al., 2004]. The disparate nature of different organisms' kinomes means that prediction programs designed for human kinases (the majority of the tools currently available) are less useful for organisms like plants. While a few plant-specific prediction tools have been developed [Heazlewood et al., 2008, Gao et al., 2009a, Durek et al., 2010], further work needs to be done both for plants and for other non-human organisms. While such work is challenging due to the smaller number of phosphorylation sites that are known for these organisms, further progress can be made as such data become more plentiful, and as structure-based methods for phosphorylation site prediction become more refined.

3.5.3 Making high-specificity predictions for whole-genome annotations

As with other classification problems, predicting phosphorylation sites involves a tradeoff between sensitivity and specificity. Greater sensitivity might be beneficial when predicting sites in a single protein, whereas greater specificity may be desirable when identifying sites in an entire proteome. This tradeoff is illustrated nicely in Table 5 of Xue et al. [2010], which shows that different tools can achieve very high specificity, but only by greatly sacrificing sensitivity (and vice versa). When sensitivity and specificity are balanced, the most accurate tools can achieve rates for both simultaneously of around 90%—a rate likely to be satisfactory when predicting sites in a limited number of proteins, but that would yield an unacceptable number of false positives when applied to an entire proteome. Unfortunately, using current prediction tools in genome annotation pipelines would therefore result in too many false positives (or too many false negatives, depending on the threshold selected). As such, the field of phosphorylation site prediction will not be truly mature until

tools are developed that offer good sensitivity combined with very high specificity.

3.5.4 Making use of evolutionary information

Many types of functional sites in proteins and nucleic acids are known to be evolutionarily conserved, such as transcription-factor binding sites [Berezikov et al., 2004], mRNA splice junctions [Shapiro and Senapathy, 1987], microRNA target sites [Friedman et al., 2009], and surface residues that participate in protein-protein interfaces [Caffrey et al., 2004]. The evolutionary conservation of phosphorylation sites has also been examined in numerous studies. For instance, the phosphorylation site Ser2 is conserved in versions of the small ubiquitin-like modifier (SUMO) protein in species as distantly related as human, *Saccharomyces cerevisiae* and *Drosophila melanogaster* [Matic et al., 2008]. Significant conservation of phosphorylation sites also occurs among different species of plants [Maathuis, 2008, Nakagami et al., 2010]. Some degree of conservation even extends to prokaryotes—although signaling via the phosphorylation of S/T/Y residues was once thought to be limited to eukaryotes, several such sites have been identified in *Escherichia coli* [Macek et al., 2008] and *Bacillus subtilis* [Macek et al., 2007].

As evolutionary information is valuable for many bioinformatics-related tasks, including protein structure prediction, gene finding, genome annotation, and sequence assembly, it should prove valuable for phosphorylation site prediction as well. For example, Jalal et al. [2009] developed a protocol that uses known human phosphorylation sites to identify putative bovine sites. While not involving machine learning, the success of this approach shows the value of using evolutionary information in order to identify novel sites. Strangely, evolutionary information has largely been ignored in the context of identifying phosphorylation sites using machine learning. In one exception, Gnad et al. [2007] used information concerning phosphorylation site conservation to improve the accuracy of their SVM-based predictor. In the future, evolutionary conservation of protein kinases (rather than, or in addition to, phosphorylation sites) may also prove useful in leveraging knowledge about one organism to predict phosphorylation sites in a second organism. Given the considerable predictive power of evolutionary information, its more widespread incorporation into future prediction tools has the potential to greatly increase accuracy.

3.6 Conclusion

There has already been a great deal of success in applying different methodologies to the problem of phosphorylation site prediction. Addressing the challenges outlined above, as well as those described by Xue et al. [2010], will require much coordination and effort, but would constitute significant steps forward for the field.

3.7 Acknowledgements

Special thanks to Chantel Krakowetz and Scott Napper for critical reading and advice, and to the reviewers for many helpful suggestions.

3.8 Funding

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Chapter 4

Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights

Brett Trost and Anthony Kusalik

This is the second of four papers that relate to the design of kinome microarrays. In this paper, a method called PHOSFER is described. PHOSFER uses a random forest-based machine-learning model in order to predict phosphorylation sites. Two primary innovations distinguish PHOSFER from the methods described in Chapter 3. First, the training data consist not only of known phosphorylation sites from the organism of interest, but also from related organisms. Each site from a related organism is weighted according to the level of phosphorylation site conservation between that organism and the organism of interest. The second innovation relates to the features used. Instead of the discrete features used by most other tools (e.g., is there a lysine residue in position 3?), PHOSFER uses a small number of real-valued features that have previously been shown to have low correlations with one another. Using soybean as a test case, it is shown that using data from other organisms results in improved accuracy compared to using only soybean data. It is also shown that PHOSFER predicts soybean sites more accurately than two tools that were designed to predict for another plant, *Arabidopsis thaliana*. Due to its use of data from organisms other than the one of interest, PHOSFER should be particularly valuable for organisms having few experimentally-characterized phosphorylation sites.

Citation

B. Trost and A. Kusalik. Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics* 29(6):686–694, 2013.

http://bioinformatics.oxfordjournals.org/content/29/6/686.long.

Author contributions

Brett Trost performed the research and wrote the paper. Anthony Kusalik supervised the research and helped edit and revise the paper.

Notes

In contrast to Section 4.6 of the manuscript, the PHOSFER web service is no longer implemented using the Galaxy platform. Instead, like the other tools available on the SAPHIRE website (http://saphire.usask.ca), custom scripts were written in order to obtain user input, run the program, and make the output available to the user.

Supplementary material

Supplementary material for this paper are given in Appendix B.

4.1 Abstract

Motivation: Phosphorylation is the most important post-translational modification in eukaryotes. While many computational phosphorylation site prediction tools exist for mammals, and a few were created specifically for *Arabidopsis thaliana*, none are currently available for other plants.

Results: In this paper, we propose a novel random forest-based method called PHOSFER (PHOsphorylation Site FindER) for applying phosphorylation data from other organisms to enhance the accuracy of predictions in a target organism. As a test case, PHOSFER is applied to phosphorylation sites in soybean, and we show that it more accurately predicts soybean sites than both the existing *Arabidopsis*-specific predictors, and a simpler machine-learning scheme that utilizes only known phosphorylation sites and non-phosphorylation sites from soybean. In addition to soybean, PHOSFER will be extended to other organisms in the near future.

4.2 Introduction

Kinase-mediated protein phosphorylation is a critical mechanism for the regulation of virtually all cellular processes in eukaryotes [Zhang and Johnson, 2000, Uddin et al., 2003, Wood et al., 2009, Wang et al., 2010, Bu et al., 2010, Ressurreição et al., 2011, Kim and Lee, 2011, Lian et al., 2010]. To fully understand signaling mechanisms in an organism of interest, it is necessary to identify both its protein kinases and the sites that those kinases phosphorylate. While the protein kinase complement of many organisms is known [e.g., Manning et al., 2002], many phosphorylation sites have yet to be identified, particularly in less well-studied organisms.

Although mass spectrometry enables phosphorylation sites to be detected in a high-throughput manner, most laboratories do not have access to the instruments and expertise required to utilize this technique. As a result, computational methods for predicting phosphorylation sites have become increasingly popular. Dozens of predictors are now available; to review these, see Xue et al. [2010] and Trost and Kusalik [2011].

Most current predictors focus on human phosphorylation sites. However, the protein kinase complements in various organisms differ significantly both in quantity and in kind [Diks et al., 2007]; for example, the plant *Arabidopsis thaliana* encodes twice as many protein kinases as does human [Champion et al., 2004], but seems to lack any that are similar to classical human tyrosine kinases. This makes most current predictors suboptimal for predicting phosphorylation sites in non-human organisms. While three tools—PhosPhAt [Heazlewood et al., 2008, Durek et al., 2010], PlantPhos [Lee et al., 2011], and an unnamed tool developed by Gao et al. [2009a]—are specific to *Arabidopsis*, predictors are lacking for other plants.

This paper describes PHOSFER (PHOsphorylation Site FindER), a phosphorylation site prediction tool for plants, particularly those for which little phosphorylation site data are available. As a test case, we use soybean (*Glycine max*), an economically important crop in many areas of the world. We utilize a novel

strategy for using phosphorylation site data from other organisms in order to boost predictive performance. Specifically, BLAST searches are used to determine the degree of conservation between phosphorylation sites in soybean and those in several other organisms for which known phosphorylation sites are available. A machine-learning scheme is employed in which a specific training instance from organism X is given a weight proportional to the level of phosphorylation site conservation between soybean and X, with greater weights implying more influence on the learning process. We show that the resultant predictors outperform the aforementioned *Arabidopsis*-specific tools when applied to soybean, and also outperform a simpler machine-learning technique that utilizes only known phosphorylation sites from soybean. In the near future, PHOSFER will be extended to predict phosphorylation sites in other organisms.

4.3 Methods

4.3.1 Data

Proteomes

The human (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*), *Caenorhabditis elegans*, and *Drosophila melanogaster* proteomes were obtained from UniProt [Apweiler et al., 2004, UniProt Consortium, 2008, 2012]. The *Arabidopsis* proteome was downloaded from The *Arabidopsis* Information Resource (TAIR) [Swarbreck et al., 2008], while the yeast (*Saccharomyces cerevisiae*) proteome was downloaded from the *Saccharomyces* Genome Database [Cherry et al., 1998, Engel et al., 2010]. Finally, the proteomes for rice (*Oryza sativa*) and soybean were retrieved from the Phytozome project [Goodstein et al., 2012].

Positive phosphorylation site data

Phosphorylation sites that have been experimentally characterized using mass spectrometry or low-throughput biological techniques were gathered from online databases for each of the nine organisms mentioned above. Known sites from *C. elegans* were gathered from Phospho.ELM [Diella et al., 2004, 2008, Dinkel et al., 2011]. Both Phospho.ELM and PhosphoSitePlus [Hornbeck et al., 2004, 2012] contained sites from human, mouse, cow, and *Drosophila*. Known sites from rice, soybean, and *Arabidopsis* were downloaded from P^3DB [Gao et al., 2009b]. Lastly, sites from *S. cerevisiae* were obtained from PhosphoGRID [Stark et al., 2010].

Although disagreement exists over the optimal peptide length for representing phosphorylation sites in a machine-learning model [Trost and Kusalik, 2011], a few studies have proposed lengths between 9 and 15 [Blom et al., 1999, Miller et al., 2008, Biswas et al., 2010] In this study, phosphorylation sites were represented as peptides of length 15, with the phosphorylated residue in the centre and seven amino acids on either side. When a particular phosphorylated residue was too close to the beginning or end of the protein to have seven residues on either side, the missing residues were represented by gap (-) characters. The handling of gaps with respect to the machine-learning features is described in Section 4.3.2. Peptides containing one

or more ambiguous amino acids were removed.

Negative phosphorylation site data

Negative phosphorylation sites (15-mer peptides with S, T, or Y central residues that are assumed not to be phosphorylated) for all organisms described earlier were gathered from their respective proteomes as follows. A given S/T/Y residue had to meet three criteria in order to be selected as a negative site. First, a potential negative site could not have been reported as a positive site. Second, as suggested by Neuberger et al. [2007], it had to be within a protein that contained known positive sites. The rationale for this criterion is that since proteins with several known phosphorylation sites have been well-studied with respect to phosphorylation, sites in these proteins that are not known to be phosphorylated are more likely to be true negatives. In this study, a potential negative site had to be in a protein containing at least three positive sites. Third, as suggested by Blom et al. [2004], a negative phosphorylation site had to be predicted as solvent-inaccessible; the rationale here is that residues buried in the core of a protein would not be accessible to any kinase. In order to predict solvent accessibility, the NetSurfP program [Petersen et al., 2009] was used. If a given S/T/Y residue was predicted as buried by NetSurfP, it was deemed to be a potential negative phosphorylation site.

Redundant sequence removal

To remove redundant sites, all positive and negative sites from all nine organisms were combined into one dataset, which was then clustered using CD-HIT [Li and Godzik, 2006] at a sequence identity threshold of 65%. These clusters were processed using the following rules.

1. If a cluster contained exactly one site, that site was retained.
2. If a cluster contained multiple positive (or negative) sites from a single organism, then a single site was arbitrarily chosen to retain.
3. Some clusters contained positive (or negative) sites from two or more organisms. To avoid redundancy, all but one of these sites were discarded. To choose the site to retain, the organism represented in the cluster with the highest level of phosphorylation site conservation with soybean (i.e., the highest value of C_{Bk}; see Section 4.3.2 for details) was determined. If there was only one site from that organism, that site was retained; otherwise, one of the sites was arbitrarily selected. Given this rule, a site from soybean was always selected if soybean was represented in the cluster.
4. Because positive and negative data from different organisms were combined, a single cluster could contain both a positive site (from one organism) and a negative site (from a different organism). If a cluster contained at least one positive site and at least one negative site, then that sequence was considered to be a positive (since the "negatives" in the other organisms are likely to be undiscovered positives). If the site was known to be a positive in more than one organism, then the organism was selected according to rule 3 above.

Dataset imbalance correction

In machine-learning problems, imbalanced datasets occur when one class has a significantly different number of instances than another class, and can significantly affect the accuracy of some learning methods [Japkowicz and Stephen, 2002]. In the context of phosphorylation site prediction, positive phosphorylation sites are vastly outnumbered by negative sites [Tang et al., 2007]. To correct this imbalance, for each organism and for each site type (S, T, or Y), the number of positive sites was determined, and an equal number of negative sites were randomly chosen from the list generated as described earlier. For example, if 123 positive sites were available for T sites in *Drosophila*, then 123 corresponding negative T sites were chosen.

4.3.2 Building the classifier

Random forests

The random forest machine-learning technique [Breiman, 2001] was used as implemented in the data mining and machine-learning package Weka [Frank et al., 2004, Witten et al., 2011]. This method involves building many decision trees, each of which is built using a number of randomly-selected features. The more trees that predict that a given peptide contains a phosphorylation site, the more likely it is that this is indeed the case. Each model built for this study utilized 300 random trees, each built using 10 randomly-selected features. Separate models were created for S, T, and Y phosphorylation sites.

Organism-specific instance weights

In this study, known phosphorylation sites both from soybean and from other organisms were used as training data. Each training instance with phosphorylated residue k ($k \in \{S, T, Y\}$) from organism B was assigned a weight based on i) the degree of phosphorylation site conservation (specifically, conservation of 15-mer peptides having phosphorylated residues in the centre) between soybean and B, and ii) the number of instances of type k in organism B. Training instances from organisms whose phosphorylation sites were better conserved in soybean were given higher weights. Conversely, the more training instances of type k that were available for a given organism, the lower the weight given to each instance. The greater the weight assigned to a particular training instance, the more influence it had on the resultant model.

Formally, let T_{Bk} represent the set of positive training instances from organism B with phosphorylated residue k. The elements of a given set T_{Bk}, as well as the negative training instances for the same B and k, were each given an identical weight W_{Bk} according to the formula $W_{Bk} = 100 \times C_{Bk}/|T_{Bk}|$. The term C_{Bk}, which represents the degree of phosphorylation site conservation between organism B and soybean, is described in more detail below. A scaling factor of 100 was applied to make the resulting numbers less unwieldy.

Each C_{Bk} was calculated as follows. Let A denote the soybean proteome, and let $(A \rightarrow B)_k$ represent the comparison in which all of the known phosphorylation sites from A (i.e., 15-mer peptides with the

phosphorylated residue in the centre) of type k were used as BLAST queries against proteome B (which could be any of the proteomes described in Section 4.3.1, including soybean itself). This was done using all the positive phosphorylation sites for a given organism, not just the ones selected at the end of the filtering process described in Section 4.3.1. Note that for phosphorylated residues occurring within 7 residues of the C- or N-terminus of a protein, the BLAST query was shorter than 15 residues, with the phosphorylated residue no longer in the middle. Specifically, let X be a known phosphorylation site from A, and let Y be its best BLAST match in B. Also, let X' and Y' denote the full-length proteins corresponding to X and Y, respectively. X was deemed to be in the "not conserved" category with respect to B if either X and Y, or X' and Y', were not conserved. X and Y were considered non-conserved if the E-value corresponding to Y was greater than 100 when X was used a BLAST query against B, or if the number of sequence differences between them was greater than or equal to 7. X' and Y' were considered non-conserved if the E-value corresponding to Y' was greater than 10^{-3} when X' was used a BLAST query against B. If X was not in the "not conserved" category according to the above criteria, then it was placed in the "conserved" category. An analogous process was also done for the comparison $(B{\rightarrow}A)_k$ (i.e., in which proteins from some proteome B were used as query sequences, and the soybean proteome was used as the database).

Let H^1_{Bk} denote the percentage of known phosphorylation sites in the "conserved" category for the comparison $(A{\rightarrow}B)_k$, and let H^2_{Bk} denote the same for $(B{\rightarrow}A)_k$. Then $C_{Bk} = (H^1_{Bk} + H^2_{Bk})/2$. For example, if 70% of sites were in the "conserved" category for $(A{\rightarrow}B)_k$ and 80% were in the "conserved" category for $(B{\rightarrow}A)_k$, then $C_{Bk} = (70 + 80)/2 = 75$. By definition, if B is soybean, then $C_{Bk} = 100$ for each k.

As an illustration of the entire step, suppose that 465 known threonine phosphorylation sites remained from rice after filtering ($|T_{Bk}| = 465$, where B is rice and k is T). Further, suppose that (prior to filtering) 27.7% of soybean T sites had conserved sites in the rice proteome, and 28.4% of rice T sites had conserved sites in the soybean proteome. Then $C_{Bk} = (27.7{+}28.4)/2 = 28.1$. The weight given to each training instance from rice (both positive and negative) would then be $W_{Bk} = 100 \times C_{Bk}/|T_{Bk}| = 100 \times 28.1/465 = 6.04$.

Features

AAIndex [Nakai et al., 1988, Kawashima et al., 2008] is a database of 544 (as of release 9.1) amino acid properties gathered from the literature. While useful for many bioinformatics tasks [Kawashima et al., 2008], the sheer number of these properties could potentially cause both computational tractability and overfitting problems when used as features in a classification problem. Given that many of these properties are strongly correlated with one another, clustering them can produce a set that is substantially smaller than the full set, but nonetheless retains much of its information. While this has been done by the authors of AAindex itself using hierarchical clustering [Tomii and Kanehisa, 1996, Kawashima et al., 2008], a more sophisticated method called consensus fuzzy clustering (CFC) was recently developed by Saha et al. [2012]. After deriving eight clusters using their technique, these authors identified a set of 24 "high-quality indices" consisting of three individual AAindex indices from each cluster: the index at the centre of the cluster (the medoid) and the

two indices farthest from the medoid. The eight clusters roughly represent electric properties, hydrophobicity, alpha and turn propensities, physicochemical properties, residue propensity, composition, beta propensity, and intrinsic propensities. A more detailed description of each of these clusters can be found in the original paper [Saha et al., 2012]; however, in order to give the reader a sense of these real-valued indices and how they relate (or do not relate) to an intuitive idea of the properties of each amino acid, Table 4.1 contains these values for three of the 24 high-quality indices. Except for the invariant middle residue, the 24 features were considered for each of the residues in a given 15-residue-long phosphorylation site, for a total of $24 \times 14 = 336$ features. As described above, missing residues (due to the phosphorylated residue being too close to the N- or C-terminus of the protein) were represented by gap characters (-) in the 15-mer representation of phosphorylation sites. In order to make it as neutral as possible, for each index the gap character was assigned a value equal to the average value for the 20 amino acids.

4.3.3 Performance evaluation

Methods compared

As described in Section 4.2, there currently exist three methods for plant-specific phosphorylation site prediction—PhosPhAt [Heazlewood et al., 2008, Durek et al., 2010], PlantPhos [Lee et al., 2011], and a method by Gao et al. [2009b]. Unfortunately, no implementation is available for the latter technique, so only PhosPhAt and PlantPhos were compared with PHOSFER. While PhosPhAt and PlantPhos were trained only using data from *Arabidopsis* (and none from soybean), they are nonetheless the two most comparable tools to PHOSFER, with other phosphorylation site tools having been trained on mammalian data [Trost and Kusalik, 2011]. (As these tools are not open-source, it was not possible to retrain them using soybean data.) Since PhosPhAt uses 13-mers rather than 15-mers (as PHOSFER does), the first and last residues were removed from each peptide before being input to PhosPhAt. PlantPhos uses 21-mers, so the 21-mer corresponding to each 15-mer site (three additional residues on either side) was used. Phosphorylated residues located too close to the beginning or end of the corresponding full protein sequence to make a full 13-mer or 21-mer could not be tested with PhosPhAt or PlantPhos, respectively. Also, PlantPhos did not return scores for a small portion of the sequences given as input, so these sites were considered to have been given a score lower than the minimum of the reported scores.

The performance of PHOSFER was also compared to those of several variants, which we have called PHOSFER-NC, PHOSFER-EW, PHOSFER-SO, PHOSFER-AO, and PHOSFER-AO25. Each was identical to PHOSFER except for the following differences. PHOSFER-NC ("no conservation") used the weights $W_{Bk} = 100/|T_{Bk}|$—that is, each training instance was weighted only according to the number of training instances for that organism, and not also according to the phosphorylation site conservation between that organism and soybean. PHOSFER-EW ("equal weights") used equal weights for all training instances, regardless of the source organism. PHOSFER-SO ("soybean only") was trained exclusively using soybean

Table 4.1: Value corresponding to each amino acid for three arbitrarily-selected high-quality indices from the clustering of amino acid properties performed by Saha et al. [2012]. Note that there are actually three indices corresponding to each of hydrophobicity, composition, and physicochemical properties; the values listed are for an arbitrarily-selected index for each. The gap character represents missing residues in the 15-mer peptides.

Amino acid	Hydrophobicity	Composition	Physicochemical properties
A	16	0.3	89.3
C	168	0.72	102.5
D	-78	1.26	114.4
E	-106	1.33	138.8
F	189	1.2	190.8
G	-13	3.09	63.8
H	50	1.33	157.5
I	151	0.45	163
K	-141	0.71	165.1
L	145	0.96	163.1
M	124	1.89	165.8
N	-74	2.73	122.4
P	-20	0.83	121.6
Q	-73	0.97	146.9
R	-70	0.9	190.3
S	-70	1.16	94.2
T	-38	0.97	119.6
V	123	0.64	138.2
W	145	1.58	226.4
Y	53	0.86	194.6
-	24	1.19	143.4

data, and thus did not involve the use of instance weights. PHOSFER-AO ("*Arabidopsis* only") was trained exclusively using *Arabidopsis* data, and thus also did not involve instance weights. Finally, PHOSFER-AO25 ("*Arabidopsis* only 25%") was the same as PHOSFER-AO, except it used only 25% of the *Arabidopsis* data for training.

Evaluating the performance of these variants allows us to assess the contribution of various aspects of PHOSFER, including the machine-learning model (random forests using AAIndex-derived features), the use of data from other species, the use of instance weights, and the use of species-specific instance weights. Specifically, comparing PHOSFER-AO with PhosPhAt and PlantPhos allowed us to compare the machine-learning model used here with those used by PhosPhAt and PlantPhos. Since there are more *Arabidopsis* data currently available than there were at the time PhosPhAt and PlantPhos were developed, we also tested PHOSFER-AO25, which used fewer *Arabidopsis* training instances than PlantPhos for each type of phosphorylation site [Lee et al., 2011] (it is not clear how many sites were used in training the PhosPhAt predictor [Heazlewood et al., 2008, Durek et al., 2010]). This allowed the impact of the machine-learning models used to be separated from the impact of a larger training set.

Comparing PHOSFER-SO with PHOSFER-AO allowed us to compare the use of soybean-specific data with the use of Arabidopsis-specific data in predicting soybean phosphorylation sites. Comparing PHOSFER-EW with PHOSFER-SO allowed us to evaluate the impact of using, in addition to soybean data, data from organisms other than soybean. Comparing PHOSFER-NC with PHOSFER-EW allowed us to evaluate the impact of weighting the training instances based on the number of instances from each organism. Finally, comparing PHOSFER with PHOSFER-NC allowed us to determine whether there is value in also weighting the training instances based on the degree of phosphorylation site conservation between soybean and the source organism.

Training and testing

Because PhosPhAt, PlantPhos, PHOSFER-AO, and PHOSFER-AO25 were trained only using data from *Arabidopsis*, they were tested directly on the known soybean data, with no cross-validation needed. PHOSFER-SO was tested using ten-fold cross-validation, with each fold using 90% of the soybean data for training and the remaining 10% for testing. PHOSFER, PHOSFER-NC, and PHOSFER-EW were evaluated in the same way, except all of the phosphorylation sites from the non-soybean organisms were used as training instances in each fold in addition to 90% of the soybean data. For completeness, in addition to ten-fold cross-validation, all performance evaluations were also done using leave-one-out cross-validation.

Evaluation criteria

For each tool, receiver operating characteristic (ROC) curves were plotted, wherein the y axis represents sensitivity, the x axis represents $1-$ specificity, and each point represents the sensitivity and specificity of a given tool at a given scoring threshold. The score is the proportion of the 300 decision trees that classify a

given residue as a phosphorylation site. Sensitivity was defined as TP / (TP + FN), where TP stands for true positives and FN for false negatives. Specificity was defined as TN / (TN + FP), where TN is true negatives and FP is false positives. Each tool was evaluated based on the area under its ROC curve (A_{ROC}), where a value of 0.5 represents classification accuracy that is only as good as random guessing, and a value of 1 represents perfect discrimination. The ROCR package [Sing et al., 2005] for the R programming language was used to facilitate the ROC analysis.

Although an A_{ROC} value represents overall classification accuracy, a classifier with a higher A_{ROC} value than another does not necessarily make it more useful. In some applications, it is important to have good sensitivity at very high specificity. For example, suppose that a user wanted to scan an entire proteome for phosphorylation sites. Because there are so many potential sites, specificity must be very high in order to avoid getting large numbers of false positives. Thus, the best tool for this situation would be the one having the highest sensitivity at very high specificity (say, 0.99). Another application might favour good sensitivity at somewhat lower specificity—for instance, specificities of 0.95 or 0.90 might be appropriate when scanning a limited number of proteins of interest. Given this, the tools were also evaluated according to their sensitivities at the practically-useful specificity values of 0.99, 0.95, and 0.90. The Matthews Correlation Coefficient (MCC) was also calculated for each of those specificity values.

4.4 Results

4.4.1 Phosphorylation site conservation and organism-specific instance weights

Positive and negative phosphorylation site data were gathered and filtered as described in Section 4.3.1. Table 4.2 shows the number of positive S, T, or Y sites from each organism (the quantities $|T_{Bk}|$ described earlier) following the removal of redundant sequences. The number of negative sites used for a given organism was made to be the same as the number of positive sites for that organism; however, due to the filtering steps performed in Section 4.3.1, cow had too few negative S sites remaining to match the number of positive sites; in this case, all possible negative sites were used. Table 4.2 also shows the level of phosphorylation site conservation between each organism and soybean; these numbers were used to calculate the values C_{Bk}. Finally, Table 4.2 contains the instance weight W_{Bk} for each combination of B and k.

4.4.2 Performance of PHOSFER, the PHOSFER variants, PhosPhAt, and PlantPhos

As mentioned earlier, the performance of the primary classifier (PHOSFER) was tested, along with those of a number of variants: PHOSFER-NC (trained using instance weights that take into account only the number of training instances from a given organism, and not phosphorylation site conservation with soybean), PHOSFER-EW (trained using equal instance weights for all organisms), PHOSFER-SO (trained using only

Table 4.2: Summary data on the known phosphorylation sites used in this study. All information is shown separately for each phosphorylated residue k (S, T, and Y) and each organism B. Column headings correspond to the notation used in Section 4.3. $|T_{Bk}|$ indicates the number of training instances that remained after filtering. H^1_{Bk} indicates the percentage of phosphorylation sites in the soybean proteome that had conserved sites in the indicated organism's proteome. H^2_{Bk} indicates the percentage of phosphorylation sites in the indicated organism's proteome that had conserved sites in the soybean proteome. Finally, W_{Bk} indicates the weight given to each individual training instance according to the formula given in Section 4.3.2. *Cow had only 92 negative S sites remaining after the filtering procedures described in Section 4.3.1, so in this case the number of negatives was less than the number of positives.

	$\lvert T_{Bk}\rvert$			H^1_{Bk}			H^2_{Bk}			W_{Bk}		
Organism	S	T	Y	S	T	Y	S	T	Y	S	T	Y
Arabidopsis	4738	1212	303	36.0%	37.5%	58.3%	30.7%	31.7%	46.1%	0.70	2.85	17.23
cow	133*	30	17	2.6%	4.9%	3.6%	1.1%	3.2%	5.3%	1.43	13.53	25.98
C. elegans	848	204	25	2.6%	2.2%	3.6%	3.3%	4.9%	17.1%	0.35	1.74	41.43
Drosophila	1107	209	36	2.2%	1.6%	2.4%	1.1%	2.8%	0.0%	0.15	1.05	3.31
soybean	332	108	49	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	30.12	92.59	204.08
human	14246	4911	7251	1.6%	3.8%	3.6%	1.7%	4.7%	7.3%	0.01	0.09	0.08
mouse	13532	2728	2207	3.3%	4.3%	3.6%	1.7%	4.3%	9.0%	0.02	0.16	0.28
rice	3396	465	118	24.1%	27.7%	35.7%	24.9%	28.4%	35.0%	0.72	6.04	29.97
yeast	3549	834	38	2.5%	4.3%	3.6%	3.4%	5.7%	11.4%	0.08	0.60	19.74

soybean phosphorylation sites), PHOSFER-AO (trained using only *Arabidopsis* sites), and PHOSFER-AO25 (training using only 25% of the available *Arabidopsis* sites). Figure 4.1 contains ROC curves illustrating the performance of PhosPhAt and PlantPhos, both of which were trained only using data from *Arabidopsis*, and compares them to PHOSFER-AO and PHOSFER-AO25. Figures 4.2 contains ROC curves for the first four PHOSFER variants mentioned above. These data were a result of using ten-fold cross-validation; results using leave-one-out cross-validation can be found in Appendix B. In addition, Table 4.3 contains the A_{ROC} value for each tool and each site type, as well as sensitivity at various practically-useful specificity values.

In Section 4.3.3, the purpose of including each of the PHOSFER variants was explained. The results given in Figure 4.1, Figure 4.2, and Table 4.3 allow the comparisons mentioned in that section to be made.

Figure 4.1 and Table 4.3 show that, for all site types, both PHOSFER-AO and PHOSFER-AO25 outperformed PhosPhAt and PlantPhos. All four of these tools were trained using data from *Arabidopsis* and then tested on soybean data, and although PHOSFER-AO had the advantage of a greater amount of training data than PlantPhos (as mentioned above, it is unclear how many training instances were used for PhosPhAt), PHOSFER-AO25 had fewer training instances than PlantPhos for all three site types. This implies that the model used here, which uses random forests and AAIndex-derived features, compares favourably with the models used by PlantPhos (and probably PhosPhAt).

As expected, Figure 4.2 and Table 4.3 show that PHOSFER-SO had the lowest performance among the variants of PHOSFER that used soybean data for training. This was the case for all three types of phosphorylation site. PHOSFER-EW, which was trained using equally-weighted data from soybean and other organisms, exhibited comparable performance to PHOSFER-SO for S sites, but greatly improved performance for T and Y sites, for which less data were available from soybean. PHOSFER-NC and PHOSFER had comparable A_{ROC} values to PHOSFER-EW for S and T sites and improved A_{ROC} values for Y sites. Finally, PHOSFER and PHOSFER-NC had comparable A_{ROC} values, but PHOSFER generally had slightly to moderately better sensitivity at high specificity than PHOSFER-NC.

Perhaps the most surprising observation from Table 4.3 is that the performance of PHOSFER-AO rivalled (and sometimes bettered) that of PHOSFER, both in terms of A_{ROC} values and in terms of sensitivity at high specificities. This observation is discussed further in Sections 4.5.3 and 4.5.4.

4.4.3 The relationship between improvements in performance and the amount of available data

Given the above observations, it appears that using data from other species provides substantial benefit when few known phosphorylation sites from the organism of interest are available (T and Y sites in this case), but a more modest benefit when many sites are available (S sites). To more explicitly examine this phenomenon, three subvariants of PHOSFER (PHOSFER75, PHOSFER50, and PHOSFER25) and PHOSFER-SO (PHOSFER-SO75, PHOSFER-SO50, and PHOSFER-SO25) were created, which used 75%, 50%, or 25%, respectively, of the available soybean data. The performance of each tool was evaluated using

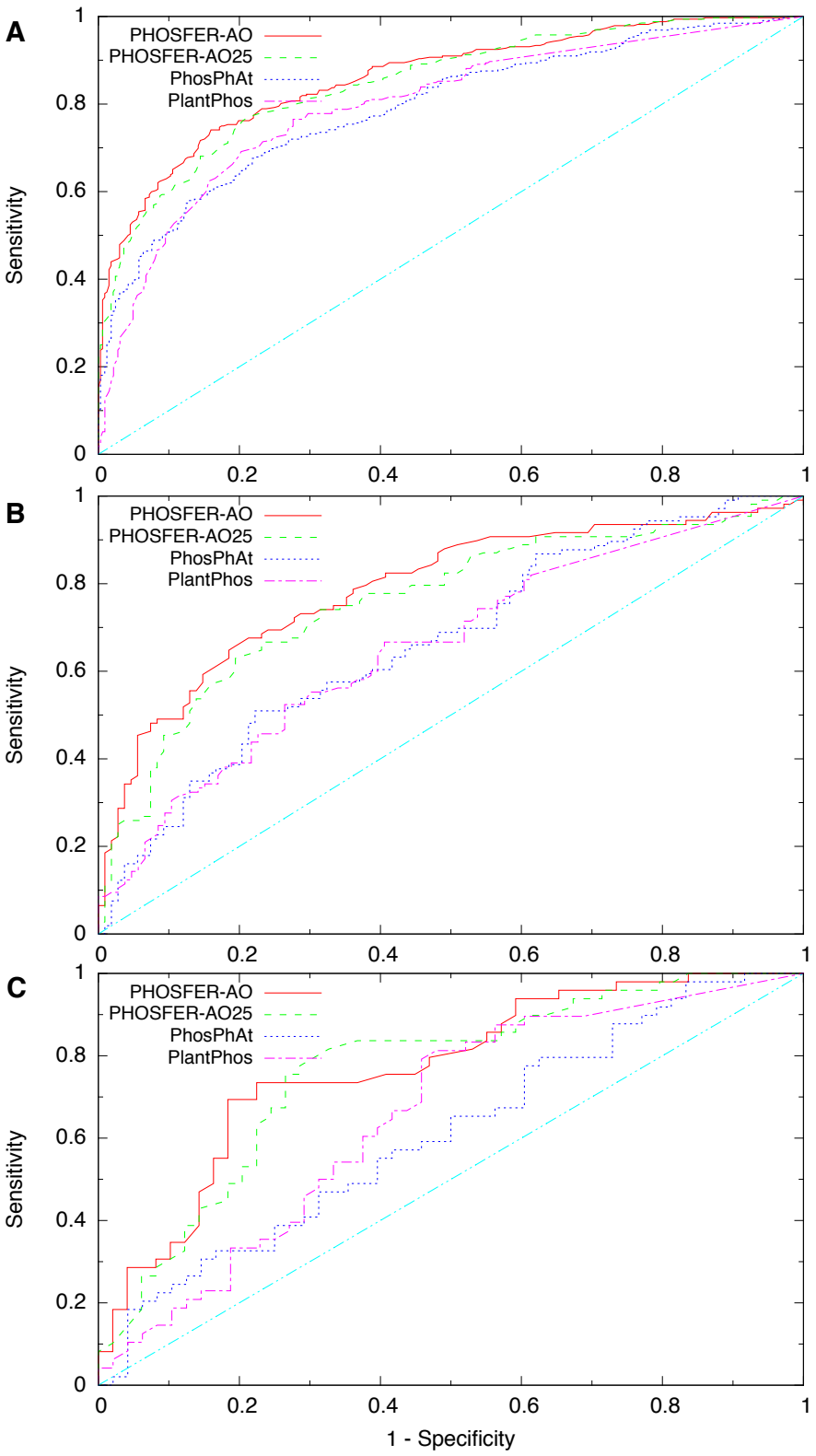


Figure 4.1: ROC curves for PHOSFER-AO, PHOSFER-AO25, PhosPhAt, and PlantPhos for (A) S phosphorylation sites, (B) T phosphorylation sites, and (C) Y phosphorylation sites. The diagonal line denotes the expected performance of a tool that uses random guessing.

ten-fold cross-validation. The results are presented in Table 4.4, which suggests that the above conjecture may be at least partially incorrect. We expected the performance of PHOSFER-SO to degrade when using smaller amounts of soybean training data, but the performance of PHOSFER to remain essentially the same; however, the performance of *both* tools remained essentially unchanged even when using only 25% of the soybean data. This may indicate that the greater improvement in performance between PHOSFER and PHOSFER-SO for T and Y sites relative to S sites cannot be attributed solely to the greater amount of soybean data available for S sites. While we cannot pinpoint with confidence an alternative explanation for this difference, it is possible that the patterns governing T and Y site recognition are more complex than those governing S site recognition, and thus benefit more from the cross-species phosphorylation site data used by PHOSFER. In any case, the fact that PHOSFER-SO, PHOSFER-SO75, PHOSFER-SO50, and PHOSFER-SO25 performed similarly shows that the machine-learning model used here (random forests using AAIndex indices as features) is robust in the face of different amounts of training data.

4.5 Discussion

4.5.1 Phosphorylation site conservation

As shown in Table 4.2, the different organisms varied greatly in the degree to which their phosphorylation sites were conserved in soybean, and vice versa. Although the numbers varied somewhat depending on the exact organism, in general the percentage of phosphorylation sites shared between soybean and another organism was about 10 times higher in the plants (*Arabidopsis* and rice) than in the non-plant organisms. For example, 24.1% of S sites in rice had a conserved site in soybean compared to just 2.6% of cow sites.

The huge disparity in phosphorylation site conservation among the different organisms means that the information provided by a training instance from one organism (e.g., human) may not be as relevant to the decision problem as one from another organism (e.g., *Arabidopsis*). This was the motivation behind the use of the C_{Bk} terms when calculating instance weights for PHOSFER.

4.5.2 Kinase specificity

While PHOSFER provides respectable accuracy for predicting phosphorylation sites in soybean, its accuracy is still less than that of most predictors that focus on human sites [Xue et al., 2010]. A portion of this underperformance could be attributed to the smaller number of known phosphorylation sites in soybean relative to human. However, a more important factor is likely the lack of information regarding the kinases responsible for phosphorylating soybean phosphorylation sites. From years of low-throughput laboratory experiments, the kinases responsible for phosphorylating many human sites are known. For example, there are currently 4,985 human sites in the PhosphoSitePlus database [Hornbeck et al., 2004, 2012] for which the corresponding kinase is known. This information allows the creation of kinase-specific predictors (the majority

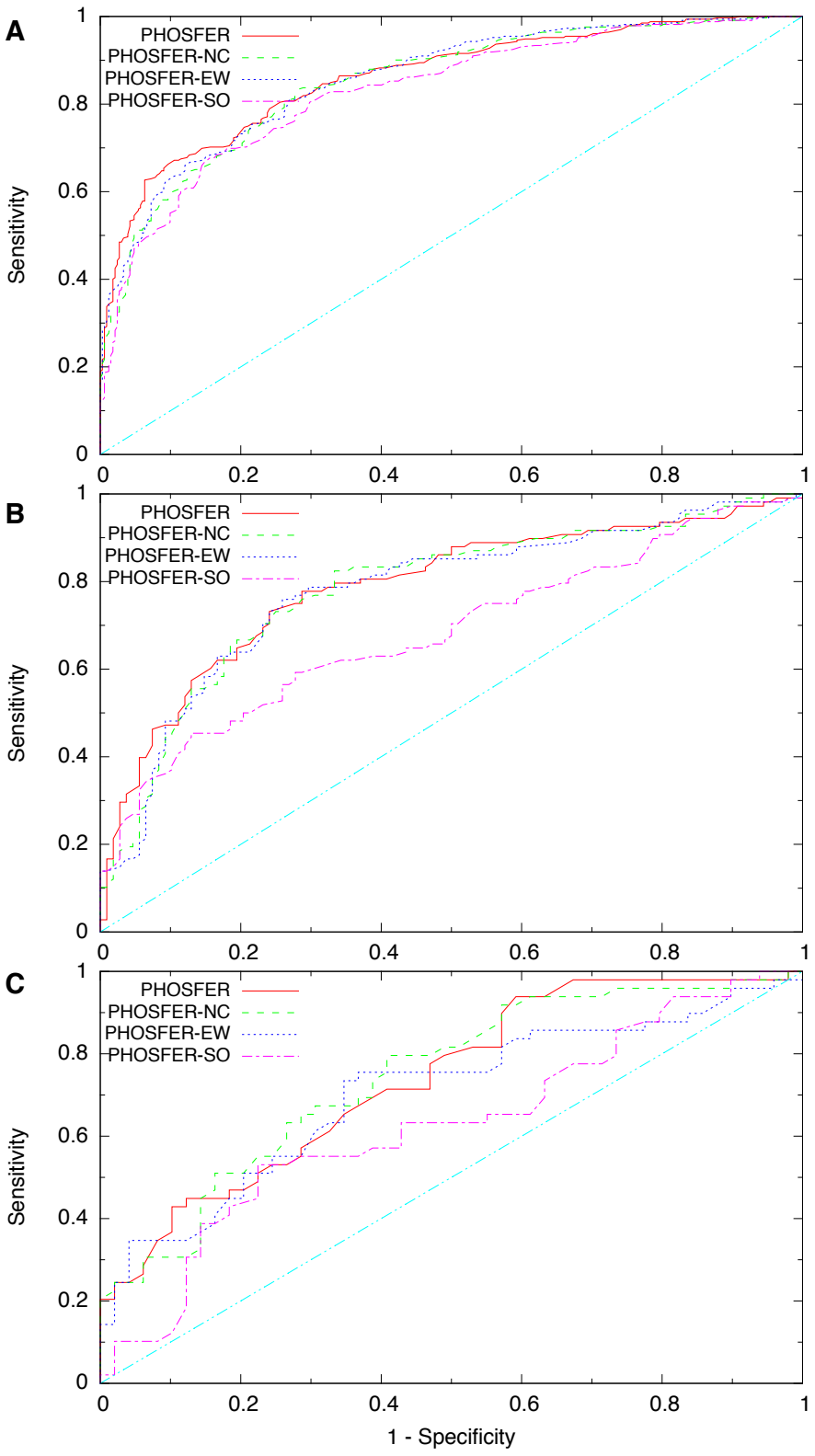


Figure 4.2: ROC curves for PHOSFER and variants for (A) S phosphorylation sites, (B) T phosphorylation sites, and (C) Y phosphorylation sites. The diagonal line denotes the expected performance of a tool that uses random guessing.

Table 4.3: Performance data for PHOSFER and its variants, as well as for the comparison tools PhosPhAt and PlantPhos. A_{ROC} values are shown, as well as sensitivity and MCC at various specificity values.

Site	Tool	A_{ROC}	Sensitivity at specificity...			MCC at specificity...		
			0.99	0.95	0.9	0.99	0.95	0.9
S	PHOSFER	0.860	0.337	0.545	0.663	0.434	0.544	0.583
	PHOSFER-NC	0.850	0.271	0.512	0.599	0.378	0.512	0.521
	PHOSFER-EW	0.857	0.307	0.482	0.630	0.409	0.491	0.551
	PHOSFER-SO	0.830	0.208	0.470	0.551	0.313	0.481	0.482
	PHOSFER-AO	0.859	0.367	0.530	0.633	0.458	0.531	0.553
	PHOSFER-AO25	0.849	0.304	0.509	0.596	0.406	0.510	0.522
	PhosPhAt	0.792	0.199	0.399	0.508	0.313	0.417	0.444
	PlantPhos	0.796	0.129	0.341	0.508	0.238	0.371	0.448
T	PHOSFER	0.788	0.167	0.324	0.472	0.278	0.358	0.409
	PHOSFER-NC	0.782	0.111	0.204	0.454	0.214	0.238	0.393
	PHOSFER-EW	0.778	0.139	0.167	0.481	0.247	0.195	0.418
	PHOSFER-SO	0.683	0.139	0.269	0.380	0.247	0.305	0.325
	PHOSFER-AO	0.789	0.185	0.352	0.491	0.297	0.383	0.414
	PHOSFER-AO25	0.760	0.111	0.259	0.454	0.214	0.296	0.393
	PhosPhAt	0.666	0.019	0.160	0.245	0.041	0.188	0.190
	PlantPhos	0.656	0.086	0.143	0.305	0.179	0.163	0.249
Y	PHOSFER	0.738	0.204	0.245	0.429	0.337	0.292	0.370
	PHOSFER-NC	0.744	0.204	0.245	0.306	0.337	0.292	0.253
	PHOSFER-EW	0.701	0.143	0.347	0.347	0.277	0.387	0.293
	PHOSFER-SO	0.624	0.020	0.102	0.122	0.102	0.119	0.032
	PHOSFER-AO	0.770	0.082	0.286	0.347	0.206	0.331	0.293
	PHOSFER-AO25	0.763	0.082	0.143	0.306	0.206	0.177	0.253
	PhosPhAt	0.609	0.000	0.184	0.245	0.000	0.224	0.185
	PlantPhos	0.655	0.042	0.104	0.188	0.146	0.120	0.118

Table 4.4: Performance comparison of PHOSFER and PHOSFER-SO when using different amounts of soybean data. PHOSFER75 and PHOSFER-SO75 were the same as PHOSFER and PHOSFER-SO, respectively, except that they used only 75% of the soybean training data; and similarly for the tools numbered 50 (50% of the soybean training data) and 25 (25%).

Tool	A_{ROC}	Sensitivity at specificity...			MCC at specificity...		
		0.99	0.95	0.9	0.99	0.95	0.9
PHOSFER75	0.870	0.305	0.566	0.671	0.409	0.561	0.586
PHOSFER-SO75	0.839	0.205	0.474	0.602	0.319	0.485	0.531
PHOSFER50	0.892	0.337	0.578	0.741	0.428	0.571	0.653
PHOSFER-SO50	0.826	0.229	0.452	0.584	0.333	0.466	0.507
PHOSFER25	0.896	0.386	0.518	0.675	0.468	0.521	0.594
PHOSFER-SO25	0.837	0.277	0.518	0.614	0.401	0.521	0.527

of current human predictors [Trost and Kusalik, 2011]), which typically have greater accuracy than non-specific predictors [Neuberger et al., 2007]. Individual kinases, as well as families of kinases, have characteristic recognition patterns, and it is likely easier to model such recognition patterns than those of kinases in general. In contrast to human, all currently known soybean phosphorylation sites were determined using mass spectrometry—a high-throughput technique that, while cheaper and faster than traditional methods of studying kinase substrates, does not provide any information on the kinases that catalyze the phosphorylation of a given site. As such, it is currently impossible to create a kinase-specific predictor for soybean, likely creating a ceiling on the accuracy of future soybean predictors—as well as predictors for any other organism for which kinase-specific information is unavailable.

4.5.3 Phosphorylation site conservation and kinase recognition patterns

It was quite surprising that although PHOSFER generally exhibited improved performance over the other PHOSFER variants that used data from soybean (PHOSFER-NC, PHOSFER-EW, and PHOSFER-SO), its performance was rivaled—and in some cases exceeded—by PHOSFER-AO. This is particularly interesting given the level of phosphorylation site conservation between *Arabidopsis* and soybean, which—while the highest of the organisms tested—was only approximately 50% for Y sites and significantly less than that for S and T sites. How, then, can using *Arabidopsis* data to predict soybean sites result in high predictive accuracy? One possibility is that, while cellular signaling pathways and processes may be only partially conserved between the two plants (thus explaining the proportion of conserved phosphorylation sites), the patterns dictating kinase recognition of those sites are more similar. If this is the case, it certainly validates the use of machine-learning tools for predicting phosphorylation sites in an organism of interest (e.g., soybean), instead of (or in addition to) simply finding conserved sites using known phosphorylation data from a related

organism (e.g., *Arabidopsis*). It would make interesting future work to statistically characterize the patterns found in the phosphorylation sites of various organisms, with statistical measures indicating their similarity or dissimilarity.

4.5.4 Testing the efficacy of simpler cross-species models

Given the comparable performance of PHOSFER and PHOSFER-AO, as additional future work it would be valuable to further investigate the performance of simpler (than PHOSFER) cross-species models. For example, would using only rice sites as training data result in similar predictive performance (relative to PHOSFER-AO) on soybean testing instances? More generally, it would be worthwhile to determine the relationship between conservation of phosphorylation sites for each organism, as shown in Table 4.2, and the efficacy of using those data as training instances for predicting soybean sites.

4.5.5 Applicability to other organisms

In addition to soybean, the technique employed by PHOSFER should be very applicable to other plants for which few known phosphorylation sites are available. For example, the number of known phosphorylation sites in economically important crops like corn, canola, and wheat, and in scientifically important model organisms like *Medicago truncatula*, are currently comparable to, or less than, that of soybean [Gao et al., 2009b, Dinkel et al., 2011, Hornbeck et al., 2012]. In addition, the technique used by PHOSFER need not be restricted to plants; the accuracy of predicting phosphorylation sites for virtually any organism of interest could be enhanced by using data from related organisms.

4.5.6 Availability

Using the Galaxy platform [Goecks et al., 2010], we have made PHOSFER available on the web. The user simply needs to browse to `http://yeoman.usask.ca` and upload a multi-FASTA file. Three tab-delimited output files will be created: one containing the score given to each 15-mer with S at its centre, and similarly for T and Y. Each file contains four columns: the name of the source protein, the 15-mer sequence, the position of that 15-mer sequence in the full protein, and the predicted score. Sequences with higher scores are more likely to be phosphorylation sites.

4.6 Conclusion

In this paper, we have described a novel machine-learning model for predicting phosphorylation sites in soybean using known phosphorylation sites from both soybean and other organisms. The use of data from other species resulted in a large improvement in predictive accuracy for T and Y sites, and a more modest improvement for S sites. The species-specific instance weights generally imparted a modest but noticeable increase in predictive performance over similar models that lacked them; however, surprisingly a model using

only *Arabidopsis* data for training was able to achieve roughly equivalent performance. We hope that the techniques outlined here—random forests, AAindex features, and the use of cross-species training data—can be used as the basis for even more accurate phosphorylation site predictors, especially for organisms having few experimentally-characterized phosphorylation sites.

4.7 Funding

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Chapter 5

DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites

Brett Trost, Ryan Arsenault, Philip Griebel, Scott Napper, and Anthony Kusalik

This is the third of four papers that relate to the design of kinome microarrays. It describes DAPPLE, which is another tool for phosphorylation site prediction. Like PHOSFER (described in Chapter 4), the primary purpose of DAPPLE is to make predictions for organisms having few experimentally-characterized sites. However, whereas PHOSFER uses a machine-learning approach, DAPPLE employs the similarity search tool BLAST (specifically, the stand-alone version from NCBI; see Section 2.3.1), with known phosphorylation sites from other organisms being used as queries, and the proteome of the organism of interest being used as the database. Compared to a previous technique that involved manual BLAST searches [Jalal et al., 2009], DAPPLE takes much less time on the part of the user, can use many more known phosphorylation sites as queries, and improves the detection of orthologues. DAPPLE is available both via a web interface (`http://saphire.usask.ca/saphire/dapple`) and via a set of scripts that users can download and run on their own machines.

Citation

B. Trost, R. Arsenault, P. Griebel, S. Napper, and A. Kusalik. DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites. *Bioinformatics* 29(13):1693-1695, 2013.

Copyright notice

Author contributions

Brett Trost helped develop the methodology, designed and wrote the software, and wrote the paper. Ryan Arsenault helped develop the methodology and provided feedback on the software. Philip Griebel helped develop the methodology. Scott Napper helped develop the methodology and provided feedback on the software. Anthony Kusalik helped develop the methodology, supervised the research, and helped edit and revise the paper.

Supplementary material

Supplementary material for this paper are given in Appendix C.

5.1 Abstract

Summary: While many experimentally-characterized phosphorylation sites exist for certain organisms, such as human, rat, and mouse, few sites are known for other organisms, hampering related research efforts. We have developed a software pipeline called DAPPLE that automates the process of using known phosphorylation sites from other organisms to identify putative sites in an organism of interest.

Availability: DAPPLE is available as a web server at `http://saphire.usask.ca`.

5.2 Introduction

Protein phosphorylation is the most widespread cellular signaling mechanism in eukaryotes [Johnson and Hunter, 2005]. Knowledge of an organism's phosphorylation sites facilitates the study of its cellular signaling pathways, which in turn has many applications in basic and translational research. Although online databases contain many phosphorylation sites for human, rat, and mouse, little data are available for other species. Using the cow as a test species, we previously proposed a protocol for making predictions in species with few known sites [Jalal et al., 2009]. Taking advantage of sequence homology between human and bovine proteins, this protocol involved manually using known human phosphorylation sites as BLAST queries to identify bovine sites. If a query and its best match in the bovine proteome had few or no sequence differences, the match was considered a putative bovine site.

While useful, several aspects of this protocol could be improved. First, its manual nature makes it time-consuming, and also limits the amount of known phosphorylation data that can be used. Second, it uses only known phosphorylation sites from human. It is possible, for instance, that a given bovine site might be homologous to a known rat site, but not to any known human site, and by using only known phosphorylation sites from human, this bovine site would be missed. This problem would be even more pronounced for species that are distantly related to human, such as plants. Third, the method used in Jalal et al. [2009] to identify non-orthologous proteins (comparing their annotations) has several drawbacks, including its subjective nature, the difficulty of automating these comparisons, and the fact that annotations are often inaccurate or incomplete.

DAPPLE is a software pipeline that addresses these concerns, ultimately allowing the user to easily, quickly, and accurately identify potential phosphorylation sites in an organism of interest.

5.3 Description of DAPPLE

A complete description of the operation of DAPPLE, including a detailed flow chart, is available as Supplementary Material (Appendix C). Below, we briefly describe the input to, and output from, DAPPLE.

DAPPLE's input files are: i) the proteome of the target organism; ii) a database of known phosphorylation

sites; and iii) the proteomes of the organisms represented in that database. All proteomes must be in FASTA format. Item iii is optional, but is necessary for DAPPLE to output information for the "RBH?" column of the output table (see below). The phosphorylation site database can be obtained from a number of sources; a partial list is included in the DAPPLE documentation. This study uses phosphorylation sites from PhosphoSitePlus [Hornbeck et al., 2012] (`http://www.phosphosite.org/downloads/Phosphorylation_site_dataset.gz`). The majority of sites in PhosphoSitePlus are represented by 15-mer peptides, with the phosphorylated residue in the middle. However, some sites are too close to the N- or C-terminus of the full protein to have 7 residues on either side, and are thus represented by a shorter peptide. To allow them to attain statistically significant BLAST hits, for these sites DAPPLE uses as a query the first or last 15 residues of the full protein sequence. As such, all queries used in DAPPLE are 15 residues in length. Additionally, entries with identical sequences (from different organisms) are removed.

The remaining phosphorylation site sequences are used as queries to `blastp`, with the target organism's proteome as the database. Unlike in Jalal et al. [2009], queries are not limited to those from human. Information about the best match (as explained in the Supplementary Materials, weaker matches may optionally be used) is saved or computed, and ultimately presented in the DAPPLE output table (described below).

Due to the short length of the query sequences, the full protein corresponding to the best match may not be orthologous to the full protein corresponding to the query. In Jalal et al. [2009], this problem was addressed by manually comparing the annotations of the two proteins. However, this approach suffers from the drawbacks described previously; thus, DAPPLE uses the well-established reciprocal BLAST hits (RBH) method to ascertain orthology [Overbeek et al., 1999]. For a known site X from organism A with match Y in target organism B, let X' be the full protein corresponding to X, and analogously for Y'. DAPPLE declares X' and Y' as orthologues if and only if Y' is the best match when X' is used as a query and the proteome of organism B is used as the database, *and* X' is the best match when Y' is used as a query sequence and the proteome of organism A is used as the database. In this case, "the best match" is defined as any protein that has the smallest E-value. Soft masking of the query sequences is used when searching full protein sequences as suggested by Moreno-Hagelsieb and Latimer [2008].

DAPPLE outputs a table in which each row represents the result of a BLAST search using, as a query, one of the known sites in the phosphorylation site database. The table is in a tab-delimited plain text format that can easily be manipulated or imported into a spreadsheet program. This table contains many columns designed to help the user decide on the accuracy and usefulness of a given match; the following list describes most of these (for the full list, see the Supplementary Materials).

- Query accession, query description, query organism, query sequence, query site—the accession number, description, organism, amino acid sequence, and phosphorylated residue (e.g., Y482) of X', respectively.
- Hit site, hit accession, hit description, hit sequence—the same as above, except for Y' rather than X'.
- Sequence differences—the number of differences between all of X (not just the portion that matched

Table 5.1: Comparison of the results of Jalal et al. [2009] with those of DAPPLE. The first column indicates the number of sequence differences between a known site from PhosphoSitePlus and its best bovine match. The second column indicates the percentage of known sites with the indicated number of sequence differences in Jalal et al. [2009]. The "no homology" row indicates known sites for which there was either no match in the bovine proteome, or the annotation of the match differed from that of the query. The third column represents output from DAPPLE, with the "no homology" row indicating that either the phosphorylation site had no match in the bovine proteome, or that "RBH?" = "no" (see Section 5.3). The fourth column is similar to the third, except instead of a site falling under the "No homology" row if "RBH?" = "no", it does so if the hit protein E-value (see Section 5.3) is greater than 10^{-5}. The E-value method represents a less stringent method of ascertaining homology (though not necessarily orthology).

Seq. differences	% (Jalal et al.)	% (RBH)	% (E-value)
0	50%	27.6%	32.9%
1	13%	14.3%	17.2%
2	7%	9.0%	11.0%
3	4%	6.2%	7.7%
4	1.5%	4.3%	5.5%
5	0.4%	3.0%	3.9%
6	0.6%	1.9%	2.6%
7+	0%	1.4%	2.0%
No homology	22%	32.2%	17.1%

in the BLAST local alignment) and Y.

- Hit protein E-value—the E-value of the match between X' and Y' when X' is used as the query and B is used as the database.

- RBH?—"yes" or "no", depending on whether X' and Y' are RBH.

5.4 Results

To test DAPPLE, phosphorylation sites in the cow (*Bos taurus*) were identified, as was done by Jalal et al. [2009]. The files described below were used as input to DAPPLE. The PhosphoSitePlus database was downloaded, and contained 214,185 unique phosphorylation sites. The proteomes corresponding to the target organism (cow) and the organisms represented in the PhosphoSitePlus database were downloaded from UniProtKB.

Table 5.1 compares the results given by Jalal *et al.* with those produced by DAPPLE. Note that both the methodology and input data used are not identical, so DAPPLE's output is not expected to be exactly the same. Nevertheless, the percentages of known phosphorylation sites that had a given number of sequence

differences with their best bovine BLAST match were similar between the two approaches. For DAPPLE, the percentage of peptides under the "no homology" category differed depending on the criterion for declaring two proteins as orthologues (see Table 5.1 caption), with the RBH method being less sensitive but more specific than the E-value method. Note that the sites reported by DAPPLE are only predictions; further, the functional significance of a homologous site may differ in the target organism, especially when the target is a distantly-related species.

Both the gain in efficiency using DAPPLE, and the value of using RBH as opposed to comparing annotations, are illustrated with examples in the Supplementary Materials.

5.5 Conclusion

DAPPLE improves upon an already-successful method for predicting phosphorylation sites for non-typical model species. Our lab has used its output to help design peptide arrays containing targets of protein kinases [Houseman et al., 2002] for studying honeybee, pig, and chicken (manuscripts in preparation), and it should be applicable to many other organisms, as well as other research problems related to protein phosphorylation. Finally, DAPPLE is not limited to phosphorylation; it could easily be applied to other post-translational modifications or to any problem that involves finding homologous motifs.

5.6 Acknowledgement

The authors thank Stephen Johnson for helping test the software.

5.7 Funding

This project was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Chapter 6

Case study: the use of DAPPLE to design a honeybee-specific kinome array

Brett Trost, Scott Napper, and Anthony Kusalik

This is the last of four papers that relate to the design of kinome microarrays. In Chapter 5, DAPPLE was illustrated by identifying phosphorylation sites in cow. Given that, like cow, most of the organisms well-represented in the phosphorylation site databases are mammals, it could be argued that cow is a relatively "easy" target organism. This chapter presents a case study in the use of DAPPLE to predict sites in honeybee (*Apis mellifera*) for the purpose of designing a honeybee-specific kinome array. This organism was chosen for several reasons: first, it is distantly related to most of the organisms that are well-represented in the phosphorylation site databases, and it is of interest to determine the usefulness of DAPPLE in such cases; second, honeybees are of substantial economic importance, as they are responsible for pollinating many crops; third, honeybee populations have recently been declining, and kinome analysis may help shed light on this phenomenon. Only the design of the honeybee-specific array is described in this chapter; the biological application of the arrays is described in Chapter 11.

The honeybee array used in the study described in Chapter 11 was originally designed in 2011 using an early version of DAPPLE whose behaviour and output were slightly different than described in Chapter 5. In addition, a significant amount of data has been added to the phosphorylation site databases since then. Therefore, the data for this chapter were derived by running the version of DAPPLE described in Chapter 5 using updated versions of the phosphorylation site databases (the same versions used to create Table 1.1).

Publication status

This manuscript has not yet been submitted for peer review.

Author contributions

Brett Trost performed the majority of the research and wrote the paper. Scott Napper performed the majority of the work in selecting peptides for inclusion on the array. Anthony Kusalik supervised the research and helped edit and revise the paper.

6.1 Abstract

Honeybees are essential pollinators for many economically and ecologically important plants. Unfortunately, worldwide honeybee populations have declined significantly in the past few years. While the causes of this decline have not yet been fully delineated, a likely contributor is the infestation of honeybees by the mite *Varroa destructor*. However, the precise physiological effects of *Varroa* infestation remain poorly understood. Given their ability to measure the phosphorylation of many targets simultaneously, kinome microarrays represent an ideal tool for studying these effects; however, the ability to design a honeybee-specific array is hampered by the absence of any known honeybee phosphorylation sites.

This chapter describes the use of DAPPLE to help design a honeybee-specific kinome array for studying the effects of *Varroa* infestation. DAPPLE was used to identify sites in the honeybee proteome that were homologous to known phosphorylation sites in the proteomes of other organisms. The identified sites were manually inspected to identify those potentially relevant to the study of *Varroa* infestation, and kinome arrays were fabricated that contained peptides representing these sites.

This case study also examined two issues relating to this process. First, the level of overlap between the contents of four phosphorylation site databases was examined, and it was discovered that the level of overlap was generally small. Second, the ability of DAPPLE to identify sites in an organism like honeybee—which is distantly related to most of the organisms represented in the phosphorylation site databases—was analyzed. It was found that although a small proportion of sites in the phosphorylation site databases had homologous sites in the honeybee proteome, the sheer number of sites in these databases (more than 200,000) meant that the quantity of predicted honeybee sites was more than sufficient for designing a honeybee-specific array. This shows that DAPPLE should be useful for identifying phosphorylation sites in (and designing kinome arrays for) organisms that are distantly related to the organisms represented in the phosphorylation site databases.

6.2 Introduction

Pollination plays a critical role in the sexual reproduction of flowering plants. While pollination can occur via wind or other abiotic factors, animals—mostly insects—perform the majority of pollination [Calderone, 2012]. Of the 107 crops that make up the vast majority of the world's agricultural output, animal-mediated pollination is absolutely essential for 13 of them, and another 57 exhibit moderate to high reductions in output in the absence of animal-mediated pollination [Klein et al., 2007].

The majority of insect pollinators are bees [Calderone, 2012]. While tens of thousands of bee species exist [Michener, 2007], some species are particularly important pollinators; for instance, the honeybee *Apis mellifera* is a critical source of pollination in many parts of the world, and was estimated to be directly responsible for $11.3 billion in crop production value in the United States alone in 2009 [Calderone, 2012].

Since approximately 2006, the number of honeybee colonies around the world has declined markedly—

a phenomenon dubbed colony collapse disorder (CCD). Given the ecological, agricultural, and economic importance of honeybees, CCD has caused a great deal of concern. As a result, substantial research efforts have been invested into identifying its causes. Several potential causes have been posited, including pesticides, viral, bacterial, or fungal diseases, nutritional deficits, poor beekeeping practices, and even increased cell phone use [Staveley et al., 2014].

One of the most likely causes for CCD is infestation by the mite *Varroa destructor* [Staveley et al., 2014]. While *Varroa* itself can harm honeybees, its most deleterious effects appear to stem from the fact that it vectors several viruses that infect bees [Rosenkranz et al., 2010]. *Varroa*-infested colonies often collapse within 2-3 years.

Despite increasing research interest, the physiological responses of honeybees to *Varroa* infestation remain poorly understood. Given that they facilitate the study of phosphorylation-mediated cellular signaling in a high-throughput manner, kinome microarrays could be a useful tool for studying the physiological effects of Varroa infestation. Unfortunately, to the authors' knowledge, there do not yet exist any experimentally characterized honeybee phosphorylation sites. Thus, designing a honeybee-specific kinome microarray requires sites to be computationally predicted. The primary objective of this study was to use DAPPLE to design an *A. mellifera*-specific kinome array that can be used to study *Varroa* infestation. Only the creation of the array is described here; a separate manuscript describing the application of the array can be found in Chapter 11.

In addition, this study had two secondary objectives. In Chapter 5, the usefulness of DAPPLE was illustrated by predicting sites in cow—an organism that, in evolutionary terms, is much more closely related than the honeybee to most of the organisms represented in the phosphorylation site databases. Therefore, one secondary objective was to determine the usefulness of DAPPLE when the target organism is distantly related to most of the organisms represented in the phosphorylation site databases, and also to determine the relative usefulness of known sites from different organisms (for example, plants versus mammals) when predicting sites in the honeybee. The other secondary objective involved comparing the contents of the phosphorylation site databases. As with many types of biological information, multiple databases dedicated to experimentally-characterized phosphorylation sites exist; however, it is not obvious how the contents of these databases relate to one another, and what the degree of overlap is among them. This study compared the contents of the four databases, which should be helpful for users wondering which databases (or combinations of databases) might be most suitable for their particular research objectives.

6.3 Methods

6.3.1 Proteomes

The proteome of *A. mellifera* was downloaded from UniProt [Apweiler et al., 2004, Boutet et al., 2007, UniProt Consortium, 2008, 2013] and contained 10,953 protein sequences. Also downloaded were the proteomes of all

organisms that had at least 10 records in one of the phosphorylation site databases.

6.3.2 Known phosphorylation sites

Data from four major phosphorylation site databases (PhosphoSitePlus [Hornbeck et al., 2004, 2012], Phospho.ELM [Diella et al., 2004, 2008, Dinkel et al., 2011], P^{3}DB [Gao et al., 2009b, Yao et al., 2012], and PhosphoGRID [Stark et al., 2010, Sadowski et al., 2013]) were downloaded on November 21, 2013. As the input to DAPPLE is expected to be in the format used by PhosphoSitePlus, the data from Phospho.ELM, P^{3}DB, and PhosphoGRID were converted into that format. Each database was then filtered in order to remove entries from any organism that did not have at least 10 entries in a single database.

Each record in a given database included a field containing the accession number of the protein in which the phosphorylation site described by that record is found. DAPPLE requires that these accession numbers match those in the corresponding proteomes. As described above, the proteomes downloaded were from UniProt; thus, in order to be used as input to DAPPLE, the phosphorylation site databases must contain UniProt accession numbers. This requirement was already met by the PhosphoSitePlus and Phospho.ELM databases, each of which provided UniProt accession numbers for each record. In contrast, records in PhosphoGRID and P^{3}DB did not necessarily have UniProt accession numbers. These databases were thus modified as follows.

The PhosphoGRID database, which contained only data from *Saccharomyces cerevisiae*, had accession numbers corresponding to the *S. cerevisiae* proteome available from the *Saccharomyces* Genome Database (SGD) [Cherry et al., 2012]. In order to construct a mapping of SGD accession numbers to UniProt accession numbers, the *S. cerevisiae* proteome was downloaded from SGD. The proteins in that proteome were used as BLAST queries against the *S. cerevisiae* proteome from UniProt. For a given SGD protein, its best match (which was nearly always 100% identical to the query protein) in the UniProt proteome was taken to be its UniProt equivalent. The data from PhosphoGRID was then modified to replace each SGD accession number with its corresponding UniProt accession number.

For P^{3}DB, the accession numbers provided with each record were inconsistent, both among different organisms and among different records from the same organism. As a result, a great deal of processing had to be performed on this dataset. The following explains what was done for each organism represented in the P^{3}DB database.

- *Medicago truncatula*—Of 15,538 records, 14,564 had accession numbers from the proteome provided by the Phytozome project [Goodstein et al., 2012]. In order to identify the Uniprot accession number corresponding to each Phytozome accession number, the Phytozome proteome was downloaded, and the proteins therein were used as BLAST queries against the UniProt *M. truncatula* proteome. The best match in the UniProt proteome for each protein from the Phytozome proteome was deemed to be its equivalent, and the P^{3}DB data were modified to replace Phytozome accession numbers with their corresponding UniProt accession numbers.

An additional 413 records had accession numbers from UniProt, and thus required no further processing. The remaining 561 records had accession numbers from various other sources; these sequences were discarded. Also discarded were records whose associated Phytozome accession numbers were not found in the current version of the *M. truncatula* proteome from the Phytozome project.

- *Arabidopsis thaliana*—Of 15,465 entries, 12,393 contained accession numbers from UniProt. The remainder contained accession numbers from The *Arabidopsis* Information Resource (TAIR) [Lamesch et al., 2012]. For these entries, the proteome from TAIR was downloaded, and a mapping between TAIR accession numbers and Uniprot accession numbers was created as previously described. A few records in P^3DB referred to accession numbers not in the current version of the TAIR proteome; these were discarded.

- *Oryza sativa* (rice)—Of 12,317 records, 9,962 contained UniProt accession numbers. The remaining records contained genetic marker loci as accession numbers. As these would be difficult to map to Uniprot accession numbers, such records were discarded.

- *Glycine max* (soybean)—All 2,739 records contained accession numbers from the soybean proteome provided by the Phytozome project, and were mapped to UniProt accession numbers as described above. Records containing Phytozome accession numbers not found in the latest version of the Phytozome *G. max* proteome were discarded.

- *Vitis vinifera* (grape)—All 862 entries had NCBI GI numbers. The Entrez batch retrieval system was used to retrieve the sequences corresponding to these GI numbers, which were used as BLAST queries against the UniProt grape proteome in order to produce a mapping of GI numbers to UniProt accession numbers. The sequences corresponding to all but one of these GI numbers had a hit against the UniProt grape proteome. The P^3DB entry corresponding to the remaining sequence was discarded.

- *Brassica napus*—The accession numbers for this organism seemingly referred to sequences available from the Computational Biology and Functional Genomics Laboratory at the Dana-Farber Cancer Institute (`http://compbio.dfci.harvard.edu`); however, the full protein sequences corresponding to these accession numbers could not be located. Therefore, all records for *B. napus* were discarded.

- *Zea mays* (corn)—All 115 entries had UniProt accession numbers.

- *Solanum tuberosum* (potato)—Of 33 entries, three different types of accession numbers were used—two from EMBL, one from UniProt, and 30 from the proteome provided by the Potato Genome Sequencing Consortium (`http://www.potatogenome.net`). Given the small amount of data and the variety of accession numbers, all potato entries were discarded.

- *Nicotiana tabacum* (tobacco)—ten entries were present containing two types of accession numbers: four from UniProt and six from The J. Craig Venter Institute (JCVI) (`http://www.jcvi.org`). For the same reasons as potato, all tobacco entries were discarded.

All databases were filtered such that if a given record contained a UniProt accession number, but that accession number was not found in the UniProt proteome downloaded for the associated organism (these accession numbers may refer to proteins that were previously present but have since been removed), then that record was removed. The data in the phosphorylation site databases were also processed to ensure that all phosphorylation sites were represented as 15-mer peptides. Where possible, each peptide was composed of the phosphorylation site at its center plus 7 residues on either side. If a given phosphorylation site was too close to the N-terminus or the C-terminus for this to be possible, the peptide was composed of the first or last 15 residues of the full protein sequence, respectively. Any record whose corresponding 15-mer contained an ambiguous amino acid was removed.

6.3.3 Examining the overlap among the phosphorylation site databases

The level of overlap in the contents of the four phosphorylation site databases was examined using the data resulting from the filtering procedures described in Section 6.3.2. This was analyzed from two perspectives: the organisms represented in each database, and—for each organism represented in more than one database—the number of phosphorylation sites from that organism that were shared or unique. Because the only non-trivial levels of overlap in the same organism occurred between the PhosphoSitePlus and Phospho.ELM databases, the latter analysis was restricted to organisms shared between these two databases. Venn diagrams were created using the R package `VennDiagram` in order to visualize the number of shared and unique sites in a given organism.

6.3.4 Examining the usefulness of known phosphorylation sites from different organisms in identifying honeybee sites

DAPPLE was run using the known phosphorylation sites that remained after the filtering steps described above, and using the proteome of *A. mellifera* as the target proteome. Some databases contained the same phosphorylation site in two or more organisms. For instance, in the PhosphoSitePlus database, the phosphorylation site S11 (corresponding to the 15-mer peptide `AAAAKKGSEQESVKE`) was found in one protein from each of human, mouse, pig, and cow (UniProt accession numbers P17612, P05132, P36887, and P00517, respectively). In Chapter 5, it was stated that DAPPLE filters its input to remove all but one of the sites in instances like this. However, here it was of interest to determine the proportion of phosphorylation sites from a given organism that had good matches in the honeybee proteome; thus, for the purposes of this study, this filtering step was not performed.

After running DAPPLE, the following summary statistics were calculated for each organism:

- the percentage of known phosphorylation sites with two or fewer sequence differences between the corresponding 15-mer peptide and its best match in the honeybee proteome ("good matches");

- the percentage of known phosphorylation sites with between three and six sequence differences between the corresponding 15-mer peptide and its best match in the honeybee proteome ("acceptable matches");
- the percentage of known phosphorylation sites with seven or more sequence differences, or no match at all, between the corresponding 15-mer peptide and its best match in the honeybee proteome ("poor matches"); and
- the percentage of sites for which the "RBH?" column was equal to "yes" (see Chapter 5).

The first three percentages add up to 100%, while the final percentage is independent of the others. If a 15-mer was found more than once in the same organism (either because it was in multiple databases, or because it was found in multiple proteins from that organism), then it was counted only once in the above calculations.

It was also of interest to determine whether the phylogenetic relatedness to honeybee of each organism represented in the phosphorylation site databases correlated with the level of phosphorylation site conservation between that organism and honeybee. To estimate the degree of phylogenetic relatedness, the complete mitochondrial genome sequence was downloaded from GenBank for honeybee, as well as for each organism represented in the phosphorylation site databases (except for the plants and yeast, whose mitochondria are not comparable to those from animals [Christensen, 2013]). The EMBOSS program `needle` was used to perform a pairwise global alignment between the honeybee mitochondrial genome sequence and the mitochondrial sequences from each organism represented in the phosphorylation site databases.

6.3.5 Identifying peptides for the honeybee-specific kinome array

The list of putative honeybee peptides generated by DAPPLE was manually inspected in order to choose those appropriate for inclusion on the honeybee-specific peptide array. Peptides were chosen to represent as wide a variety of signaling pathways and metabolic processes as possible, but with an emphasis on proteins involved in stress responses or innate immunity. Other criteria used in selecting peptides included the number of sequence differences between the 15-mer peptide corresponding to the known phosphorylation site and its best hit in the honeybee proteome (with fewer sequence differences being preferred), the value of the "RBH?" column (with "yes" being preferred), and the location of the phosphorylation site in the honeybee protein relative to its location in the query protein (with similar locations being preferred to more distant locations).

6.4 Results

6.4.1 Proteomes

As described in Section 6.3.1, the proteome of *A. mellifera*, as well as the proteomes of organisms having at least 10 records in one of the phosphorylation site databases, were downloaded from UniProt. The first

column of Table 6.1 contains a list of these organisms.

6.4.2 Known phosphorylation sites

Table 6.1 summarizes the number of records remaining from each organism in the four databases after performing the procedures described in Section 6.3.2 (in contrast, Table 1.1 gives the number of records prior to filtering). In some cases, the number of post-filtering sites was equal to, or only slightly less than, the original number of sites contained in the database. For instance, no sites from PhosphoGRID were removed during filtering (Tables 1.1 and 6.1), and only 12 of 862 sites were removed for grape. However, for some organisms, a substantial number of sites were removed during filtering; for instance, the number of sites from rice was 12,317 before filtering but only 7,850 after filtering.

6.4.3 Examining the overlap among the phosphorylation site databases

Given that there are multiple phosphorylation site databases, a user of DAPPLE (or anyone interested in known phosphorylation site data) may wonder how the contents of those databases compare. In this study, four databases were compared both at the organism level (that is, the number of phosphorylation sites from a given organism that were present in each database) and at the sequence level (if multiple databases each had sites from the same organism, to what degree did those sites overlap?).

Table 6.1 shows that the level of overlap at the organism level among the four phosphorylation site databases was quite small. For instance, after filtering, PhosphoGRID contained 6,440 phosphorylation sites from *S. cerevisiae* (the only organism represented in this database), while Phospho.ELM—the only other database containing entries from *S. cerevisiae*—contained just 57. With two small exceptions (a single site from soybean and three sites from corn in Phospho.ELM), the organisms represented in P^3DB were not represented in any of the other databases. The only two databases that had a significant amount of overlap between them were PhosphoSitePlus and Phospho.ELM. Of the 18 organisms that were represented in at least one of these databases, nine were represented in both, while six were represented only in Phospho.ELM and three were represented only in PhosphoSitePlus. For most of the organisms represented in both databases, PhosphoSitePlus contained more sites—often substantially more—than Phospho.ELM. For instance, PhosphoSitePlus contained over 150,000 human sites versus fewer than 36,000 in Phospho.ELM. The ratios of mouse and rat sites were even more biased in favour of PhosphoSitePlus—68,062 versus 7,255 and 9,358 versus 544, respectively. With respect to the six organisms that were present in Phospho.ELM but not PhosphoSitePlus, four (yeast, soybean, corn, and pacific herring) had only a few sites; however, Phospho.ELM contained 2,278 sites from *Drosophila melanogaster* and 1,470 from *Caenorhabditis elegans*—organisms absent from every other database.

Given that both PhosphoSitePlus and Phospho.ELM contained sites from many of the same organisms, it was of interest to determine the degree to which the phosphorylation sites from a given organism in PhosphoSitePlus overlapped with those in Phospho.ELM. For each of the nine organisms represented in both

Table 6.1: Number of phosphorylation sites for each organism in each major phosphorylation site database after filtering using the procedures described in Section 6.3.2.

Organism	PhosphoSitePlus	Phospho.ELM	P^3DB	PhosphoGRID
Homo sapiens (human)	150,612	35,425	0	0
Mus musculus (mouse)	68,062	7255	0	0
Rattus norvegicus (rat)	9,358	544	0	0
Medicago truncatula	0	0	13,515	0
Arabidopsis thaliana	0	0	14,791	0
Oryza sativa (rice)	0	0	7,850	0
Saccharomyces cerevisiae (yeast)	0	57	0	6,440
Caenorhabditis elegans	0	1470	0	0
Drosophila melanogaster (fruit fly)	0	2278	0	0
Glycine max (soybean)	0	1	2,092	0
Vitis vinifera (grape)	0	0	850	0
Brassica napus (rapeseed)	0	0	0	0
Bos taurus (cow)	463	188	0	0
Gallus gallus (chicken)	334	102	0	0
Oryctolagus cuniculus (rabbit)	169	89	0	0
Sus scrofa (pig)	80	18	0	0
Zea mays (corn)	0	3	107	0
Xenopus laevi (frog)	33	0	0	0
Mesocricetus auratus (hamster)	0	0	0	0
Canis lupus familiaris (dog)	40	5	0	0
Solanum tuberosum (potato)	0	0	0	0
Ovis aries (sheep)	11	12	0	0
Torpedo californica (pacific electric ray)	2	0	0	0
Clupea pallasii (pacific herring)	0	10	0	0
Capra aegagrus hircus (goat)	9	0	0	0
Nicotiana tabacum (tobacco)	0	0	0	0

PhosphoSitePlus and Phospho.ELM, the number of sites found only in the former database, only in the latter database, or in both databases were determined. Venn diagrams representing these results are given in Figure 6.1, which shows that, for most organisms, the majority of the sites present in Phospho.ELM were also present in PhosphoSitePlus. However, Phospho.ELM had some unique sites in eight of the nine organisms that were represented in both databases.

6.4.4 Examining the usefulness of known phosphorylation sites from different organisms in identifying honeybee sites

DAPPLE was used to identify potential matches in the honeybee proteome for each phosphorylation site listed in Table 6.1. Table 6.2 summarizes the percentage of phosphorylation sites from a given organism whose corresponding 15-mer had a given number of sequence differences with its best match in the honeybee proteome. For all organisms represented in the phosphorylation site databases, the percentage of their phosphorylation sites having good matches in the honeybee proteome was small. For instance, just 3.9% of human sites had a match with two or fewer sequence differences in the honeybee proteome, and 16.4% had a match with between three and six sequence differences. These percentages were similar for most other mammals, such as mouse and rat. Pig and sheep were slight exceptions, with higher percentages of good matches than the other mammals. However, this may be an anomaly resulting from the small number of phosphorylation sites available from these organisms (92 and 17, respectively). As expected, the plants had consistently lower percentages of good matches compared to the mammals; for instance, just 10.4% of sites from *Arabidopsis thaliana* had matches in the honeybee proteome with six or fewer sequence differences. Unexpectedly, fruit fly—the most closely related organism to honeybee of the organisms in Table 6.2—was little or no better than the mammals in terms of phosphorylation site conservation with honeybee.

The results described above are quite different from those in Chapter 5, where the bovine proteome was the target. In that case, approximately 60% of the sites in the PhosphoSitePlus database had a match with two or fewer sequence differences in the bovine proteome—a reflection of the fact that cow is much more closely related to the organisms represented in the phosphorylation site databases than is honeybee.

To determine whether the evolutionary relatedness between a given organism and honeybee correlated with the level of phosphorylation site conservation between them, the mitochondrial genome sequence was downloaded for each organism, and the percent identity between a given organism's mitochondrial genome sequence and that of the honeybee was determined. These values are given in the last column of Table 6.2. The correlation between the percentage of phosphorylation sites in a given organism having 6 or few sequences differences with the best match in the honeybee proteome (i.e., the sum of the second and third columns in Table 6.2) and the percent identity of the mitochondrial genome sequences was determined. Interestingly, the correlation between the two variables was actually negative ($r = -0.22$). However, when pacific herring and goat (which had anomalous phosphorylation site conservation values due to the small number of known sites from these organisms) were excluded, then the correlation was close to zero ($r = -0.06$). This lack of

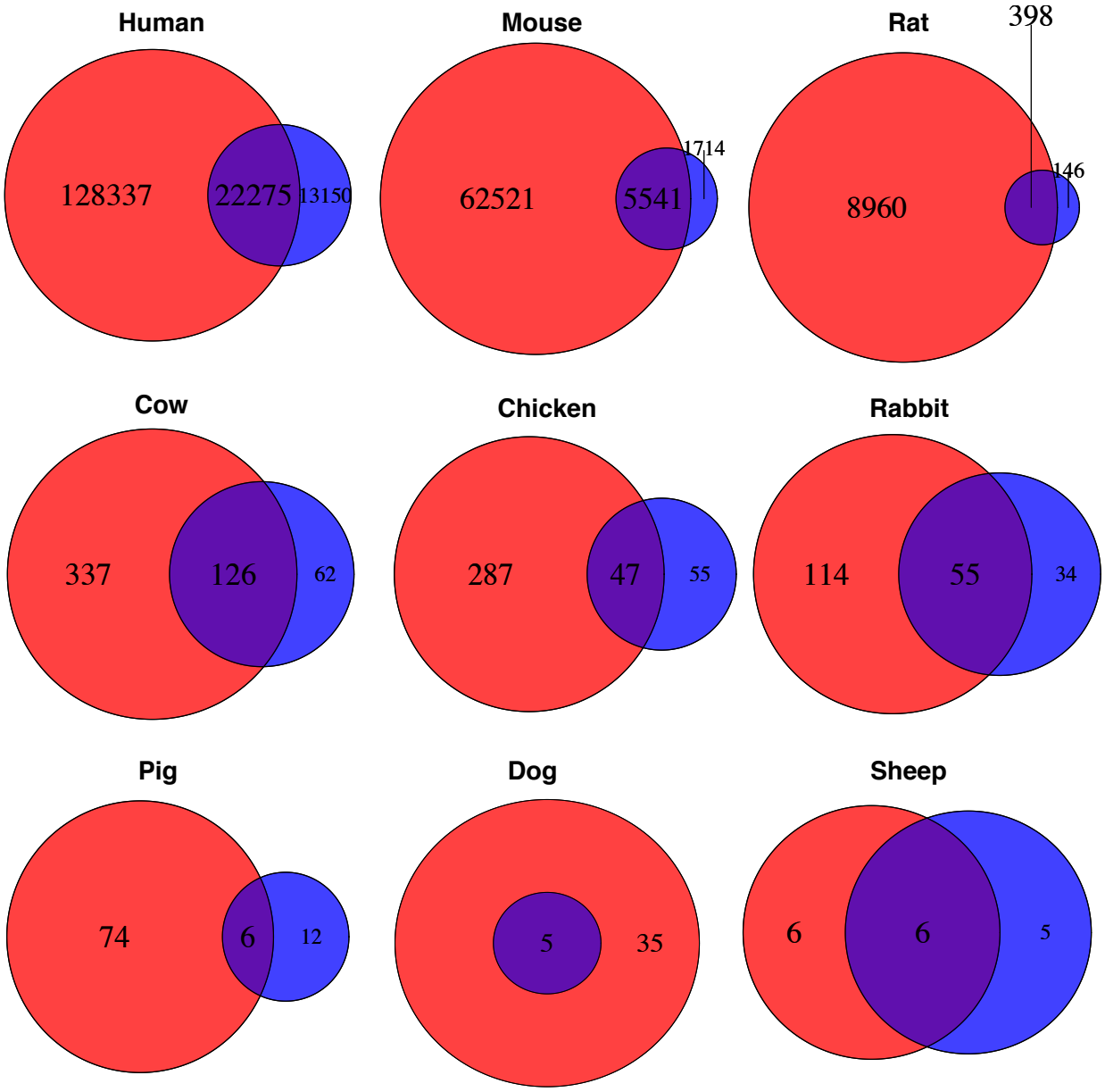


Figure 6.1: The number of phosphorylation sites found in PhosphoSitePlus only (red), Phospho.ELM only (blue), or both databases (purple) for each of the nine organisms that were represented in both databases. The sets are drawn to scale within each organism, but not between organisms.

correlation is unlikely to be the case in general; its presence in this study can likely be explained by the fact that all of the organisms represented in the phosphorylation site databases are quite distantly related to the target organism (honeybee).

6.4.5 Identifying peptides for the honeybee-specific kinome array

As mentioned earlier, the main goal of this study was to use DAPPLE to identify potential peptides for inclusion on a honeybee-specific kinome array. A total of 201,913 unique 15-mers (each of which contained a known phosphorylation site) were used as input to DAPPLE. Of these, 35,411 had matches with six or fewer sequence differences in the honeybee proteome, and 5,995 had matches with two or fewer sequence differences. From these sequences, 299 peptides were chosen for inclusion on the honeybee array. While it would not be practical to explain the rationale for choosing all of these peptides, the following list provides some details on three of the honeybee peptides that were selected.

- `DLDHERMSYLLYQML`—The central residue in this peptide corresponds to residue S135 in the honeybee protein with UniProt accession number H9KH67. The known phosphorylation site that was used to identify this peptide was S129 in the mouse protein with accession number Q91Y86 (annotated as "mitogen-activated protein kinase 8"), which corresponds to the 15-mer peptide `ELDHERMSYLLYQML`. There were a number of reasons why `DLDHERMSYLLYQML` was chosen. First, there was only one sequence difference between this peptide and its counterpart in the mouse protein. Second, the proteins with accession numbers H9KH67 and Q91Y86 were reciprocal BLAST hits (that is, the "RBH?" column in the DAPPLE output was "yes"). Third, the location of the phosphorylation site in the mouse protein (residue 129) was close to its location in the honeybee protein (residue 135), giving further evidence of the correspondence between the two sites. Finally, although the honeybee protein is annotated only as "uncharacterized protein", it appears to be an orthologue of mitogen-activated protein kinase 8, a protein known to be involved in a diverse array of stress responses. Therefore, characterizing the phosphorylation of this protein would be relevant when studying *Varroa* infestation.

- `HKLGGGQYGEVYEGV`—The central residue of this peptide corresponds to residue Y300 in honeybee protein H9K2C5. The phosphorylation site Y253 in the human protein with accession number P00519 (which corresponds to the 15-mer `HKLGGGQYGEVYEGV`) was used to identify this peptide. Although the honeybee protein was again annotated only as "uncharacterized protein", the human protein—with which the honeybee protein appears to be orthologous—is annotated as "tyrosine-protein kinase ABL1", another protein associated with stress responses. This peptide was chosen for reasons similar to those given for the previous peptide: the residue locations were reasonably similar (300 versus 253), and the number of sequence differences between the two peptides was small (2).

- `YKERIDEYDYAKPLE`—The central residue of this peptide corresponds to residue Y1510 in honeybee protein H9KFK6. It was identified via the known phosphorylation site Y596 in the rat protein with

Table 6.2: Degree of phosphorylation site conservation between *A. mellifera* and each organism represented in the phosphorylation site databases. The second, third, and fourth columns list the percentage of known phosphorylation sites from the organism in the first column that had 0–2 sequence differences, 3–6 differences, or 7 or more differences, respectively, between a given 15-mer sequence and its best match in the honeybee proteome. The "RBH?" column lists the percentage of sites for which the "RBH?" column of the DAPPLE output table was equal to "yes" (see Chapter 5). The final column lists the percent identity between the sequence of the mitochondrial (mt) genome from that organism and the sequence of the honeybee mitochondrial genome. Values were not determined for cells marked "N/A"; this was either because the organism's mitochondrial genome was not comparable to that of honeybee (for plants and yeast), or because the complete mitochondrial genome sequence was not available (for *Xenopus laevi* and *Torpedo californica*).

Organism	# sequence differences			RBH? (%)	mt % identity
	0–2 (%)	3–6 (%)	7+ (%)		
Homo sapiens (human)	3.9	16.4	79.7	9.4	32.8
Mus musculus (mouse)	3.0	14.5	82.5	7.5	37.0
Rattus norvegicus (rat)	5.5	16.7	77.9	11.1	37.1
Medicago truncatula	1.1	10.3	88.7	1.8	N/A
Arabidopsis thaliana	0.4	10.0	89.5	0.9	N/A
Oryza sativa (rice)	0.8	9.2	90.1	0.9	N/A
Saccharomyces cerevisiae (yeast)	1.1	11.5	87.4	3.3	N/A
Caenorhabditis elegans	2.9	15.5	81.6	7.4	47.2
Drosophila melanogaster (fruit fly)	5.0	18.4	76.6	13.1	52.8
Glycine max (soybean)	0.9	10.9	88.2	1.8	N/A
Vitis vinifera (grape)	0.5	11.8	87.7	1.9	N/A
Bos taurus (cow)	4.6	13.0	82.4	8.8	35.0
Gallus gallus (chicken)	5.2	18.9	75.9	13.6	30.4
Oryctolagus cuniculus (rabbit)	7.2	13.8	79.0	12.2	37.7
Sus scrofa (pig)	12.8	12.8	74.4	8.1	32.7
Zea mays (corn)	0.0	9.8	90.2	3.9	N/A
Xenopus laevi (frog)	6.1	3.0	90.9	0.0	N/A
Canis lupus familiaris (dog)	0.0	12.8	87.2	2.6	37.0
Ovis aries (sheep)	11.8	23.5	64.7	0.0	37.1
Torpedo californica (pacific electric ray)	0.0	0.0	100.0	0.0	N/A
Clupea pallasii (pacific herring)	0.0	100.0	0.0	0.0	32.8
Capra aegagrus hircus (goat)	0.0	0.0	100.0	0.0	37.3

accession number Q920L2, which corresponds to the 15-mer `YKVRIDEYDYSKPIE`. Compared to the two honeybee peptides described above, this was a more speculative choice. First, the phosphorylated residues were far apart (residue 1510 versus residue 596); second, the proteins were not reciprocal BLAST hits; third, the level of similarity between the honeybee peptide and the rat peptide was slightly lower (3 sequence differences). The rat protein is annotated as "succinate dehydrogenase", an enzyme involved in energy production in mitochondria.

Once the list of honeybee peptides had been selected, kinome arrays were fabricated by a commercial partner (JPT Peptide Technologies, Berlin, Germany; `http://www.jpt.com`).

6.5 Discussion

6.5.1 Examining the overlap among the phosphorylation site databases

For a given type of biological data, it is often the case that two or more databases store data of that type. For instance, general sequence data are present in databases maintained by both NCBI and the European Bioinformatics Institute (EBI). As another example, there are at least three databases dedicated to 16S rRNA sequences: the Ribosomal Database Project [Maidak et al., 1997, Cole et al., 2014], SILVA [Pruesse et al., 2007, Quast et al., 2013], and GreenGenes [DeSantis et al., 2006]. Phosphorylation sites represent yet another type of data for which there are multiple sources, with several databases acting as repositories for this type of information (four of which were examined here). While multiple databases can sometimes confer benefits to the user, such as a broader spectrum of tools for data analysis, it can also produce confusion (for example, it may not be clear how the data in the different databases relate) as well as frustration (for example, different databases usually present their data in different formats, requiring the user to convert data from one format to another in order to combine data from different databases).

As mentioned in the introduction, one of the secondary goals of this chapter was to compare the contents of four major phosphorylation site databases. The results presented in Section 6.4.3 suggest that none of the four databases are rendered completely redundant by any of the others. PhosphoGRID and P^3DB were particularly unique, with very few of their sites being present in any other database. The degree of overlap between PhosphoSitePlus and Phospho.ELM was greater, with many sites being present in both databases. In cases where the two databases each contained data from a given organism, PhosphoSitePlus usually had far more sites than Phospho.ELM; however, for most such organisms Phospho.ELM contained a non-trivial number of sites that were not found in PhosphoSitePlus.

Given the information in Table 6.1 and Figure 6.1, it appears that the most appropriate database to use for DAPPLE depends on the organism in which phosphorylation sites are being predicted. Clearly, if the target organism is a plant, then P^3DB would be the most appropriate database; similarly, if the target organism is closely related to yeast, then PhosphoGRID should be chosen. If the target organism is a mammal,

PhosphoSitePlus would be preferable over Phospho.ELM due to its greater number of sites; however, there is nothing precluding a user from also using data from Phospho.ELM in order to increase the number of known sites. If the organism of interest is closely related to *Caenorhabditis elegans* or fruit fly, then Phospho.ELM ought to be used, as it is the only database containing sites from these two organisms. Finally, there is no apparent disadvantage to using all four databases in order to use the largest possible amount of data.

6.5.2 Examining the usefulness of known phosphorylation sites from different organisms in identifying honeybee sites

In Section 6.4.4, it was reported that only a small percentage of the known phosphorylation sites in the phosphorylation site databases had good matches in the honeybee proteome. However, given the sheer number of sites contained in these databases, DAPPLE was still able to identify more than enough putative honeybee phosphorylation sites to allow the design of a honeybee-specific array. As reported in Section 6.4.5, nearly 6,000 known phosphorylation sites had good (two or fewer sequence differences) matches in the honeybee proteome, and nearly 30,000 additional sites had moderately good matches (between three and six sequence differences). Given that kinome microarrays typically contain a few hundred unique peptides, these numbers should be more than adequate for the purposes of kinome microarray design. Thus, in addition to being a useful tool for identifying phosphorylation sites in organisms that are closely related to those represented in the phosphorylation site databases (as was shown for cow in Chapter 5), DAPPLE also appears to be useful for identifying sites in more distantly-related organisms.

In identifying phosphorylation sites in an organism like honeybee, a user might be tempted to use as input to DAPPLE only sites from fruit fly, which is the most closely related organism to honeybee among the organisms represented in the phosphorylation site databases. However, there are at least two reasons why this may not be a good strategy. First, although fruit fly may be more closely related to honeybee than the other organisms, it is not closely related to it in an absolute sense. While fruit fly and honeybee belong to the same class (Insecta) in the taxonomic hierarchy, they belong to different orders (Diptera and Hymenoptera, respectively); also, the fact that their mitochondrial DNA sequences are only 53% identical (Table 6.2) further supports the argument that these two organisms are not closely related. Second, the data in Table 6.2 suggest that the level of phosphorylation site conservation between fruit fly and honeybee is no better than between the mammals and honeybee. Therefore, if a user was to use as input to DAPPLE only known sites from fruit fly, it may not result in a sufficient number of potential honeybee sites in order to design an array. Specifically, if it is assumed that a site in the honeybee proteome with six or fewer sequence differences from its corresponding known phosphorylation site is a candidate for inclusion on an array, then running DAPPLE using only known phosphorylation sites from fruit fly would give approximately 500 potential honeybee sites. While this is greater than the approximately 300 unique sites that are typically included on an array, it does not provide a great deal of choice in selecting specific peptides from pathways of interest.

6.6 Conclusion

This chapter described the successful application of DAPPLE to identify putative phosphorylation sites in the honeybee proteome. These sites were manually inspected in order to find those potentially relevant to the study of honeybee infestation by the mite *Varroa destructor*. Kinome microarrays containing the peptides corresponding to each of the selected sites were then fabricated. The results of applying these arrays to *Varroa*-infested honeybees can be found in Chapter 11.

In addition to the selection of peptides for a honeybee-specific kinome array, this study also compared the contents of four major phosphorylation site databases, which should be useful both for users of DAPPLE and for those interested in known phosphorylation sites in general. Finally, while Chapter 5 showed that DAPPLE is useful for identifying phosphorylation sites in organisms that are closely related to the organisms represented in the phosphorylation site databases (such as cow), this chapter showed that it can also be used to identify sites in more distantly related organisms (like honeybee).

Chapter 7

A systematic approach for analysis of peptide array kinome data

Yue Li, Ryan J. Arsenault, Brett Trost, Jillian Slind, Philip J. Griebel, Scott Napper, and Anthony Kusalik

This is the first of two papers that relate to the analysis of kinome microarray data. Previously, researchers using kinome arrays would use software designed for DNA arrays in order to analyze their data. In this paper, a software pipeline called PIIKA is described that is specifically designed for the analysis of data from kinome arrays. It contains features for background subtraction, transformation and normalization, detection of peptides that have inconsistent phosphorylation signals among technical or biological replicates, statistical comparisons among samples, and clustering. Using a case study involving molecules known to induce specific signaling pathways, it is shown that PIIKA is better able to identify differentially modulated signaling pathways than other methods.

Citation

Y. Li, R. J. Arsenault, B. Trost, J. Slind, P. J. Griebel, S. Napper, and A. Kusalik. A systematic approach for analysis of peptide array kinome data. *Sci Signal* 5(220):pl2, 2012.

Author contributions

The original problem was posed by Scott Napper and Philip Griebel. Yue Li initially devised the pipeline with input from Ryan Arsenault and Scott Napper and completed the first implementation. The implementation was subsequently revised and refined by Brett Trost and Jillian Slind. The case study was devised by Ryan Arsenault, Scott Napper, and Philip Griebel, with lab work performed by Ryan Arsenault and data analysis and comparison by Yue Li. The first form of this manuscript was composed by Yue Li. Extensive revisions were performed by Anthony Kusalik and Brett Trost. The work was supervised by Scott Napper, Anthony Kusalik, and Philip Griebel.

Notes

After this manuscript was published, the web address for PIIKA was changed. Rather than being located at `http://www.cs.usask.ca/research/research_groups/combi/piika`, as stated in the manuscript, the website can be accessed via `http://saphire.usask.ca/saphire/piika`. However, the former web address redirects to the latter address, so those reading this paper will still be able to access the correct site.

Supplementary material

Supplementary material for this paper are given in Appendix D.

7.1 Abstract

The central roles of kinases in cellular processes and diseases make them highly attractive as indicators of biological responses and as therapeutic targets. Peptide arrays are emerging as an important means of characterizing kinome activity. Currently, the computational tools used to perform high-throughput kinome analyses are not specifically tailored to the nature of the data, which hinders extraction of biological information and overall progress in the field. We have developed a method for kinome analysis, which is implemented as a software pipeline in the R environment. Components and parameters were chosen to address the technical and biological characteristics of kinome microarrays. We performed comparative analysis of kinome data sets that corresponded to stimulation of immune cells with ligands of well-defined signaling pathways: bovine monocytes treated with interferon-γ (IFN-γ), CpG-containing nucleotides, or lipopolysaccharide (LPS). The data sets for each of the treatments were analyzed with our methodology as well as with three other commonly used approaches. The methods were evaluated on the basis of statistical confidence of calculated values with respect to technical and biological variability, and the statistical confidence (P-values) by which the known signaling pathways could be independently identified by the pathway analysis of InnateDB (a Web-based resource for innate immunity interactions and pathways). By considering the particular attributes of kinome data, we found that our approach identified more of the peptides involved in the pathways than did the other compared methods and that it did so at a much higher degree of statistical confidence.

7.2 Introduction

Eukaryotic protein kinases constitute a large and important superfamily of enzymes. There are over 500 members that catalyze approximately 100,000 unique phosphorylation events in humans [Manning et al., 2002, Zhang et al., 2002]. Functionally, kinases are at the core of signal transduction and have central roles in virtually every cellular behavior, including metabolism, transcription, cytoskeletal rearrangement, and immune defense. The central roles of kinases in regulating cellular processes and diseases, as well as their conserved catalytic cleft, make them logical targets for drug therapy. In addition, there is a growing appreciation that investigations of cellular responses at the level of phosphorylation-mediated signal transduction offer considerable insights into phenotypes and mechanisms of cellular responses.

The experimental approaches for the analysis of cellular phosphorylation can be divided into analyses of the kinome and the phosphoproteome depending on whether the focus of the analysis is the protein kinases that mediate phosphorylation (the kinome) or the protein targets of these kinases (the phosphoproteome). The most substantial challenges to phosphoproteome analysis are the low abundance of phosphoproteins relative to other proteins within the proteome, and that many proteins are phosphorylated to substoichiometric extents, so that only a small fraction ($\sim$1%) is modified at any given time [Mann et al., 2002]. Another limitation of phosphoproteome analysis is that it is often conducted with phosphorylation-specific antibodies,

which are of limited availability. A promising alternative to phosphoproteome analysis is to focus on the kinome, because the well-defined, highly conserved chemistry of enzymatic phosphorylation enables rapid characterization of kinase activity, provided that appropriate substrates are available.

Proteins are the physiological substrates for most kinases. Because the specificities of most kinases are dictated by the amino acid residues that surround the phosphorylation site, a logical alternative to the study of whole proteins is to use substrate peptides that represent these sequences. These peptides can be excellent kinase substrates, with V_{max} (the maximum rate at which an enzyme can catalyze a reaction) and K_m (the amount of substrate required for the enzyme to function at one half of its maximal rate) values that are close to those of the natural substrates [Kemp et al., 1977]. Peptides are easily produced, relatively inexpensive, chemically stable, and highly amenable to array technology. To date, most peptide arrays that are generated for kinome analysis have been based on phosphorylation events characterized from a particular species and are used for analysis of that same species. However, because phosphorylation sites and their biological consequences are often conserved, we hypothesized that it would be possible to predict the sequence contexts of phosphorylation events in proteins of other species on the basis of genomic information. In 2009, we used this bioinformatics approach to develop an array of 300 bovine peptides, each of which had high sequence conservation to a human peptide containing a known phosphorylation site [Jalal et al., 2009].

Through the adaptation of high-throughput microarray technologies originally developed for gene expression analysis, it is possible to explore kinase activity in a given species [Parikh and Peppelenbosch, 2010]. However, kinome microarray experiments have several features that are distinct from those of typical gene expression experiments. First, the number of peptides on a kinome microarray is approximately two orders of magnitude smaller than the number of oligonucleotides or complementary DNAs (cDNAs) embedded on a transcription array [Jalal et al., 2009, Fletcher et al., 2009]. Thus, it is not desirable to discard data points because they are deemed "outliers" or because they are negative values (which cause problems for a typical log transformation). In addition, peptides may be recognized by the correct protein kinase, but with a lower efficiency than occurs when the sequence is in the context of an intact protein [Jalal et al., 2009]. Moreover, kinome activities may vary across individual subjects within the same species (for example, between different human patients). Thus, the reduced but still existing problem of dimensionality (that is, the number of variables is much greater than the number of samples) and the distinct biological nature of the data may make unsuitable the approaches commonly practiced in gene expression analysis [Smyth, 2004, Fundel et al., 2008a,b, Fletcher et al., 2009]. This unsuitability primarily concerns rigorously testing for the variability between biological replicates, imposing statistical stringency on the differential analysis, and recognizing signaling pathways from the differential phosphorylation information obtained.

Some kinome studies have exploited a standard and commonly practiced approach in gene expression analysis [Löwenberg et al., 2006, Fundel et al., 2008a,b, Jalal et al., 2009]). Briefly, after background correction, intensities in each kinome array are normalized to the 50th or 90th percentile of the data points from the same array. Any peptide with a SD that is larger than 1.96 times the mean of its replicate data points from

the same array is deemed an "outlier" and is removed from further analyses. The average is taken over the replicate spots for each of the remaining peptides. The fold-change ratio for each peptide under a treatment is calculated relative to the control [van Baal et al., 2006, Löwenberg et al., 2006, Jalal et al., 2009, Arsenault et al., 2009]. Fold-change ratios above or below a certain threshold are considered statistically significant for the phosphorylated or dephosphorylated peptides, respectively.

Limitations of this approach lie within the context of weakly phosphorylated peptides. For example, the statistical significance of a (de)phosphorylation fold-change ratio of 1.5 is higher in the context of high-intensity readings than in a low-intensity range. Furthermore, correction for background intensities may result in the generation of negative values, for which fold-change is nonapplicable and ratios are meaningless. These latter data points are either set to an arbitrary value or are removed from further analysis. Unfortunately, both strategies discard potentially useful data [Huber et al., 2002]. Linear Models for Microarray Data (*limma*), a popular software package at Bioconductor (`http://www.bioconductor.org`), facilitates normalization of data generated from cDNA microarray experiments and analysis of differential expression for multi-factor design experiments [Wettenhall and Smyth, 2004]. Applications of *limma* for kinome analyses are emerging. For example, in the study of chondrosarcoma by Schrage et al. [2009], *limma* is applied after quantile normalization to the kinome data set consisting of 1024 different kinase substrates in triplicate with 16 negative and 16 positive controls. The resulting moderated t-statistics appear to underestimate the true significance of the kinome data, and very few phosphorylated substrates have adjusted P-values less than 0.05, which is a commonly accepted measure of statistical significance. This reflects the need to treat data from kinome analyses differently from those of transcription profiles [Smyth, 2004]. In this case, a less stringent statistical inference method is desirable.

We have established a software pipeline for kinome analysis that addresses the aforementioned challenges (Figure 7.1). For ease of reference, the pipeline is called PIIKA, an acronym for "Platform for Intelligent, Integrated Kinome Analysis." PIIKA is primarily implemented as a script in the programming language R [R Development Core Team, 2006]. A prototype deployment of the pipeline as a Web-based server and corresponding graphical user interface has also been completed. PIIKA was designed with the goal of identifying truly differentially phosphorylated peptides specific to a treatment under investigation, while eliminating misleading factors that interfere with the interpretation of results. A set of statistical procedures was chosen to address the problems of variability that exist among technical and biological replicates. Visualization based on statistical significance is also provided. The identifiers of the differentially phosphorylated peptides can be used to search for known signaling pathways from reliable resources such as InnateDB (`http://www.innatedb.ca`) [Lynn et al., 2008] or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (`http://www.genome.jp/kegg`) [Kanehisa and Goto, 2000, Kanehisa et al., 2006, 2010]. The results may elucidate the pathways specifically stimulated by the treatment under study, thus providing insight into the mechanisms that particular cells use in response to the tested stimuli. Last, we incorporated hierarchical clustering and principal component analysis (PCA) into the methodology for comparative visualization of

phosphorylation patterns under various treatments. PCA is capable of reducing the number of variables to only the two or three most important ones (that is, the principal components) that account for most of the variability in the data sets. The data points corresponding to the samples can then be plotted with the derived components in order to examine their clustering pattern.

To demonstrate its utility, we applied PIIKA to the analysis of phosphorylation data sets from treatment conditions in which the signaling pathways are known and documented. The pipeline's outputs were compared with those from three alternate methodologies that were applied to the same input data. The compared methodologies are previously described techniques for the analysis of microarray data, two of which have been applied to peptide microarrays. The four methods were compared on the basis of their abilities to reflect the known biology with statistical confidence. A Web site has been established (`http://www.cs.usask.ca/research/research_groups/combi/piika`) to make PIIKA, and information about PIIKA, available to the research community. At the site are sample data from the CpG and LPS treatment experiments described in the Case Study under Notes and Remarks, as well as usage documentation (similar to that available in the Instructions). The source code for PIIKA (usage of which is described in the Instructions) will be available under a no-charge license solely for academic, noncommercial research in the near future. Instructions for obtaining a license are at the Web site.

7.3 Materials

The data that are entered into our new method of kinome analysis are raw signal intensities from kinome microarrays in the form of tables in tab-delimited text files. Each text file represents one physical kinome microarray, with the rows of the table corresponding to spots on the array. The exception is the first row, which contains column headings. The columns of the table are the name of the peptide, the accession number of the protein containing the peptide, the foreground intensity measurement, the background intensity measurement, and designation of phosphorylation site. The initial 10 rows of a file conforming to the prescribed format are given in Table 7.1.

Replicates within the array (for example, multiple spots for individual peptides on a single array) must be listed one after another in the table. Thus, if a particular array has 300 distinct peptides, each of which is repeated three times, then the corresponding file would have 901 lines, including the line that contains the column headings. It is assumed that there are at least two conditions (and therefore at least two files), one corresponding to a biological control and the other to a specific treatment. In addition to replicate peptides within arrays (intraarray replicates), there may also be multiple arrays for the same subject and treatment (interarray replicates) or biological replicates (for example, different subjects). However, for simplicity, there cannot be both interarray replicates and biological replicates. Also, it is assumed that the numbers of intra-array replicates per array, interarray replicates per treatment, and subjects per treatment are all constant.

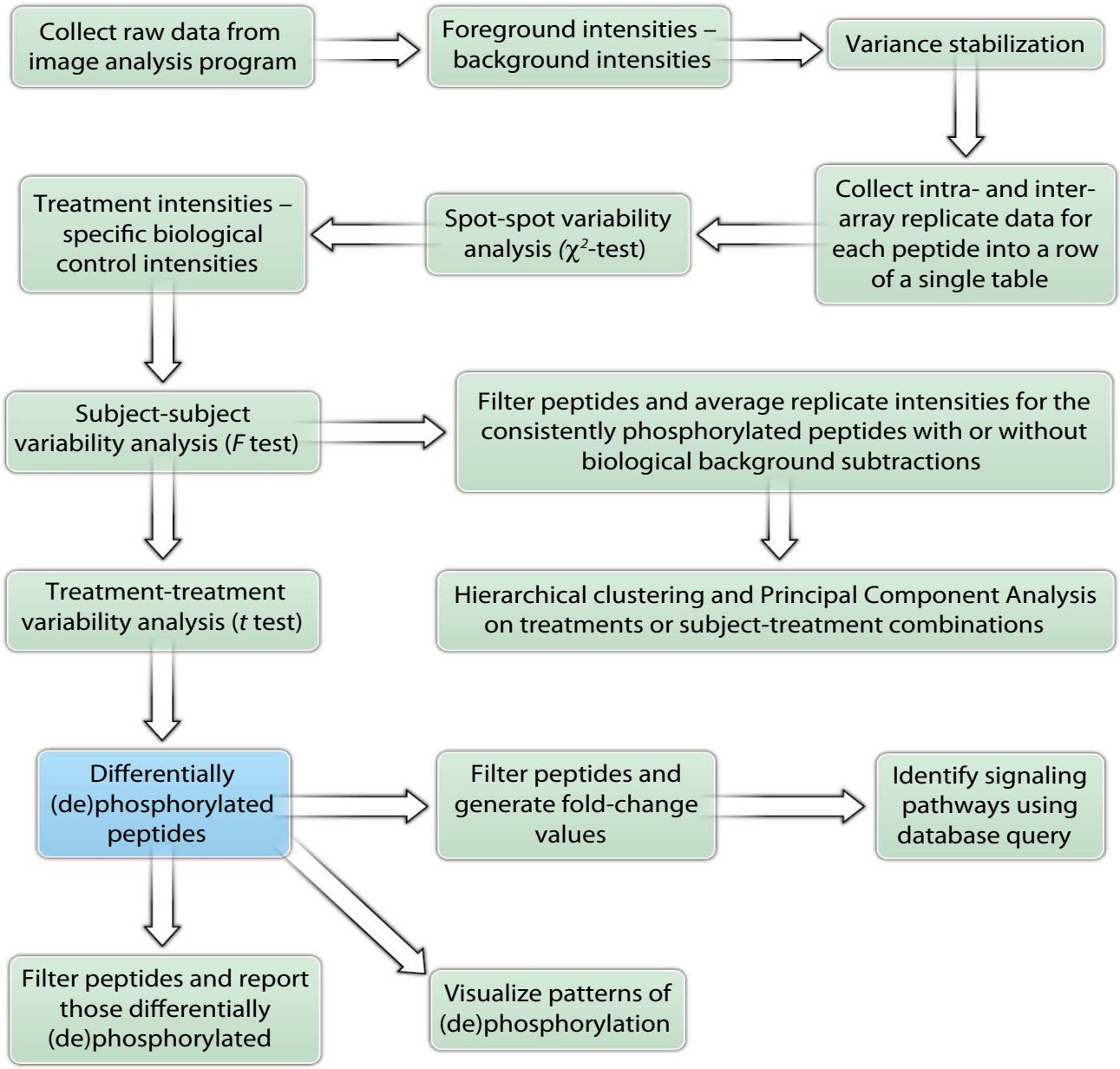


Figure 7.1: A general workflow of the proposed method for kinome analysis. The flow of information and activity starts from the top left and follows the arrows. Rectangles with green background represent procedures, and the one with blue background represents intermediate results.

Table 7.1: The initial 10 rows of a sample file conforming to the prescribed format described in the Materials. Name, name of the protein containing the peptide; ID, accession number of that protein; F532 Mean, foreground intensity measurement as provided by the image analysis software; B532 Mean, background intensity measurement; Target, the residue in the intact protein that is phosphorylated.

Name	ID	F532 Mean	B532 Mean	Target
4E-BP1	Q13541	16920	16707	T46
4E-BP1	Q13541	17213	16869	T46
4E-BP1	Q13541	17962	17436	T46
4E-BP1	Q13541	18364	17588	T37
4E-BP1	Q13541	6013	5657	T37
4E-BP1	Q13541	6025	5357	T37
A-Raf	P10398	15850	14187	Y301
A-Raf	P10398	18867	17747	Y301
A-Raf	P10398	19844	17995	Y301

7.4 Equipment

A computer with R [R Development Core Team, 2006], bash (or sh), and Perl software installed under UNIX, Mac OS X, or LINUX is necessary. PIIKA works with R version 2.11.1 running on Mac OS X version 10.6.7 and on a 64-bit Mandriva LINUX Distribution 10.0, kernel version 2.6.31.14. Subsequent versions of R should be upwardly compatible. No special features of bash or Perl are used. Other necessary software includes the *vsn* (version 1.12.0), *scatterplot3d* (version 0.3-33), and *gplots* (version 2.8.0) packages for R. The *scatterplot3d* and *gplots* packages are available from `http://cran.r-project.org/web/packages`, whereas the *vsn* package can be obtained at `http://www.bioconductor.org`.

7.5 Instructions

These instructions provide a step-by-step guide to setting up and running PIIKA on Mac OS X, UNIX, or LINUX. They also include a running example using the data from experiments involving treatment of cells with CpG or LPS (see the Case Study under Related Techniques) and available from `http://www.cs.usask.ca/research/research_groups/combi/piika`. For the purposes of this Protocol, the word "experiment" refers to a set of treatments that are to be compared for their effect on phosphorylation. Thus, all of the data from the CpG and LPS treatments and their controls are considered to be from one experiment. Except for step 1, the procedure described here would be performed separately for each experiment. In examples involving UNIX or LINUX shell commands, the $ symbol represents the shell command prompt and thus should not be typed. Descriptive names in angle brackets (<>) are to be replaced by names conforming to the

descriptor in the angle brackets. Occasionally, these instructions refer briefly to specific features of the PIIKA methodology. Further information on these is given in the PIIKA Methodology section in the Supplementary Materials, which provides a very detailed set of instructions for performing the analysis independently of the provided software.

7.5.1 Downloading PIIKA

1. Download the scripts.

   Note: The scripts are made available as a compressed "tar" saveset named "piika.tar.gz". See the PIIKA Web site (`http://www.cs.usask.ca/research/research_groups/combi/piika`) for information on obtaining this file.

2. After downloading this file, open a virtual terminal program in which a UNIX or LINUX shell is running and change the current working directory of the shell to the directory in which you saved the file "piika.tar.gz."

3. Next, use the command "$ tar -xzf piika.tar.gz" to decompress and unpack the file.

   Note: This will create a directory called "PIIKA" that has a single subdirectory called "scripts." The "scripts" directory contains the main PIIKA script ("piika.R"), as well as two accessory scripts, the shell script "init.sh" and the Perl script "create_combined_file.pl."

7.5.2 Running PIIKA

1. Change your current working directory to the PIIKA directory. If you just performed the instructions described in "Downloading PIIKA," then type "$ cd PIIKA." Otherwise, use the appropriate command involving "cd" to change your current working directory to the "PIIKA" subdirectory that was created when you unpacked the file "piika.tar.gz."

2. Use the provided script "init.sh" to create a directory corresponding to your experiment. The usage of this script is as follows: "$ scripts/init.sh <name of experiment>".

   Note: Underscores, rather than spaces, should be used in naming your experiment. The script "init.sh" creates a new directory called "<name of experiment>", which has four subdirectories: "data," "config," "<name of experiment>_results," and "intermediate_results." Each of these subdirectories should be empty except for "<name of experiment>_results," which has an empty subdirectory called "t-tests." For example, you can use the following command to create a directory for the CpG and LPS experiment: "$ scripts/init.sh CpG_LPS"

3. Change your current working directory to the data subdirectory. Use the command "$ cd <name of experiment>/data" to change your current working directory to the newly created data subdirectory. For the CpG/LPS experiment, use "$ cd CpG_LPS/data."

4. Establish data files. Place the raw data files from your experiment in the current working directory. Each file should contain the data from a single microarray and have the format described in the

Materials. To continue with the CpG and LPS example, perform the following steps to place the data files from the experiment in the "data" subdirectory. Download the example data from the PIIKA Web site (`http://www.cs.usask.ca/research/research_groups/combi/piika`) and save it as "cpg_lps.tar.gz" in the current working directory. Then, use the command "$ tar -xzf cpg_lps.tar.gz" to decompress and unpack the "cpg_lps.tar.gz" file.

5. Create a file containing the data from all arrays. In this step, the "create_combined_files.pl" script is used to create a file containing the data from all of your arrays in the experiment. This "combined file" is used as input to "piika.R." The usage of "create_combined_files.pl" is

   "$ ../../scripts/create_combined_file.pl <file 1> <file 2> ... <file N>"

   where N is the number of files.

   Note: In our running example, $N = 4$, and the following command would be issued (all on one line): "$ ../../scripts/create_combined_file.pl CpG_treatment.txt CpG_control.txt LPS_treatment.txt LPS_control.txt." In the CpG and LPS example, there are no biological replicates or interarray replicates. Thus, the files could have been specified in any order. If biological or technical replicates were present, then the files must be grouped by treatment—that is, all of the files corresponding to the first treatment, followed by all of the files corresponding to the second treatment, and so on. Two files are created: "combined.txt" is the main input file to "piika.R" (which will be used in step 8), whereas "file_numbers.txt" indicates the order in which the columns occur in "combined.txt." Although "file_numbers.txt" is not used as input to PIIKA, it may be helpful for performing the next two steps.

6. Specify pairs of treatments for comparison. In this step, you specify which pairs of treatments you want to compare. For each treatment pair specified, a t-test will be performed for that pair, and biological subtraction will be performed for that pair to generate a heatmap and for PCA analysis (Supplementary Materials, PIIKA Methodology). To complete this step, first change your current working directory to the "config" subdirectory created in step 2 by issuing the command: "$ cd ../config." Then, in this subdirectory, create a text file called "t-tests.txt," each line of which specifies a pair of treatments—the first treatment, then a tab, and then the second treatment (for example, the control).

   Note: Each treatment is specified by a number defined by the content of "combined.txt." The first treatment in "combined.txt" is 1, the second treatment is 2, and so on. If there are no biological or interarray replicates, then the order of the treatments corresponds exactly to the contents of "file_numbers.txt." Otherwise, "file_numbers.txt" can still be used to remind the user of the order of the treatments, but the fact that there is more than one file per treatment must be taken into account.

   Note: Continuing the CpG/LPS example, the "t-tests.txt" file should look like this:

   1 2
   3 4

   This tells PIIKA to perform a t-test between "CpG_treatment.txt" and "CpG_control.txt" (treatments 1 and 2), and another t-test between "LPS_treatment.txt" and "LPS_control.txt" (treatments 3 and 4). Because of the specification in "t-tests.txt," PIIKA will also perform biological subtraction on both pairs of treatments to generate a heatmap and a PCA plot.

7. Specify treatment combinations for visualization. In a fashion similar to the previous step, now specify the treatments that should be compared with the visualization method described in PIIKA Methodology (Supplementary Materials). To do this, create a text file called "visualizations.txt" in the current working directory. This file has the same format as "t-tests.txt," except that each line has four numbers separated by tabs. The first two numbers specify the two treatments to be compared in the left semicircle, and the last two numbers specify the two treatments to be compared in the right semicircle. For example, to compare the CpG treatment with the CpG control in the left semicircle, and the LPS treatment against the LPS control in the right semicircle, "visualizations.txt" should have the following contents:

1 2 3 4

Note: For a given semicircle, that pair of treatments must also appear in the file "t-tests.txt." Thus, the following "visualizations.txt" file would not be valid (given the contents of the "t-tests.txt" file used earlier), because there is no line in "t-tests.txt" that contains 3 in the first column and 1 in the second column:

3 1 3 4

8. Run PIIKA. To run PIIKA, change your current working directory to the directory called "<name of experiment>" that you created in step 2. If you are currently in the "<name of experiment>/config" subdirectory, you can accomplish this with the following command: "$ cd ..".

   PIIKA is executed as follows. Because of space constraints, the following is shown on multiple lines, but the actual command should consist of just a single line:

   "R --no-restore --no-save --args <input data filename> <number of intra-array replicates> <number of treatments> <number of replicate arrays> <number of unique peptides> <replicate type> <do χ^2-test?> <do F-test?> <do biological subtraction before doing F-test?> < ../scripts/piika.R"

   Note: The parameters (program arguments) <number of intra-array replicates>, <number of treatments>, <number of replicate arrays>, and <number of unique peptides> are self-explanatory. Descriptions of the remaining parts of this command are as follows. The "--no-restore", "--no-save", and "--args" options are necessary to run R noninteractively and must be given as specified. The "<input data filename>" parameter is the path to the input data file. This is typically a file created in step 5. The "<replicate type>" parameter must be either "biological" or "technical," depending on whether the replicate arrays (if any) represent biological or technical replicates. If the number of replicate arrays is 1, then the value of this parameter is irrelevant (but must still be specified). The way in which the χ^2-test is performed differs depending on which option is chosen. If the "technical" option is chosen, then all of the replicates (the number of intraarray replicates multiplied by the number of replicate arrays) for a given peptide and treatment combination are used in the χ^2-test to determine whether the responses for that peptide or treatment are inconsistent. If the "biological" option is chosen, then a χ^2-test is performed separately on each replicate array, and a peptide-treatment combination is considered inconsistent if the P-value for any of the arrays corresponding to that treatment is less than the P-value threshold (0.01). The final three parameters must be either "yes" or "no" depending on whether you want to perform the χ^2-test, perform the F-test, or do biological subtraction before performing the F-test, respectively. The final part of

the command ("< ../scripts/piika.R") specifies that R is to read and execute the instructions in the named file (that is, the PIIKA implementation script).

Note: In the CpG and LPS example, the input data are in the file "data/combined.txt," the number of intraarray replicates is 3, the number of treatments is 4, the number of replicate arrays is 1, and the number of unique peptides is 300. There is only one replicate array for each treatment; thus, it does not matter whether we choose "biological" or "technical" for the replicate type. Also, we want to perform the χ^2-test, but we do not want the F-test (indeed, we cannot perform it) because we do not have multiple subjects for each treatment. Because we do not perform the F-test, the final parameter is unimportant, so we can enter "no." The full command would therefore be (all given on a single line): "$ R --no-restore --no-save --args data/combined.txt 3 4 1 300 technical yes no no < ../scripts/piika.R".

Note: Assuming that the script completes successfully, there should be a number of files created in the "<name of experiment>_results" subdirectory ("CpG_LPS_results" in the running example). These files are as follows: "PCA.pdf" contains both the two-dimensional (2D) and 3D PCAs. These are graphical images in PDF format; "PCA_biological_subtraction.pdf" contains the 2D and 3D PCAs after biological subtraction (between each pair of treatments in the file "t-tests.txt") is performed. These are graphical images in PDF format; the "heatmaps.pdf" file contains hierarchical clustering results.

Note: There should be three images in the "heatmaps.pdf" file, each of which shows the clustering with different distance metrics and linkage methods. The first heatmap uses average linkage and Pearson correlation. The second heatmap uses complete linkage and Euclidean distance, whereas the third uses McQuitty linkage and Pearson correlation. The file "heatmaps_biological_subtractions.pdf" contains hierarchical clustering results after biological subtraction (between each pair of treatments in the file "t-tests.txt") is performed. As before, there are three images in PDF format. The file "visualization_<name of visualizaton>.pdf" contains visualization files as specified in "visualizations.txt." The name of each file depends on the names of the treatments in that particular visualization. Each file contains one image in PDF format. In the "<name of experiment>_results/t-tests" subdirectory, the results of each t-test specified in "t-tests.txt" are given. For each line of "t-tests.txt," three files are generated. The "<name of test>.all.txt" file includes all peptides. The "<name of test>.significant.txt" file includes all peptides for which either the P-value for an increase in extent of phosphorylation or the P-value for decreased phosphorylation is less than 0.1. The file "<name of test>.consistent.txt" includes all peptides that are found to be consistently phosphorylated through both the χ^2-test (for technical replicates) and the F-test (for biological replicates).

9. To analyze another data set, repeat the instructions beginning at step 1.

7.6 Related techniques

Some kinome studies have used a standard and commonly practiced approach from gene expression analysis [Löwenberg et al., 2006, Fundel et al., 2008a,b, Jalal et al., 2009]. That approach has already been outlined in the Introduction. Also popular is *limma* (Linear Models for Microarray Data), one of the most popular Bioconductor packages in R (`http://www.bioconductor.org`). It provides functions for the normalization of cDNA microarray data and the analysis of differential expression for multifactor design experiments [Wettenhall and Smyth, 2004]. The differential analysis component of the *limma* package uses an empirical Bayes (eBayes) model that estimates the SEs in the expression of each gene by borrowing information across genes

and calculating the moderated t-statistic accordingly [Smyth, 2004]. The limitations of both approaches were described earlier. Further detail on related techniques, as well as a comparison of the performance of PIIKA with those of the other techniques, is given in the following case study.

7.6.1 Case study

To demonstrate the viability of PIIKA, we applied it to phosphorylation data sets from experiments in which the signaling pathways are known and documented. The experiments involved exposure of bovine monocytes to interferon-γ (IFN-γ), oligonucleotides containing CpG motifs [which are well-characterized Toll-like receptor 9 (TLR9) ligands], and lipopolysaccharide (LPS), which stimulates TLR4. IFN-γ is responsible for activating macrophages to clear intracellular pathogens [Dorman and Holland, 1998, Döffinger et al., 2000]. Signal transduction by IFN-γ is associated with a specific Janus kinase (JAK)-signal transducer and activator of transcription (STAT) signaling cascade [Darnell, 1997, Pestka et al., 1997]. The microbe-associated ligand CpG activates pathways involving TLRs, which are pathogen recognition receptors that alert the host to the presence of microbial challenge [Arsenault et al., 2009]. Lastly, treatment of immune cells with LPS induces an increase in the abundance of the interleukin-2 (IL-2) receptor [Le et al., 1986, Scheibenbogen et al., 1992], which leads to increased IL-2-dependent signaling. To establish the value of PIIKA, we compared its outputs with those from three alternate methodologies applied to the same input data. The compared methodologies are previously described techniques for the analysis of microarray data, two of which have been applied to peptide microarrays. The compared methods are percentile normalization (PNorm) + fold-change (FC) [Löwenberg et al., 2006, van Baal et al., 2006, Arsenault et al., 2009, Jalal et al., 2009], quantile normalization (QNorm) + *limma* [Schrage et al., 2009], and VSN + *limma* [Fletcher et al., 2009]. We compared the four methods on the basis of their abilities to reflect the known biology with statistical confidence.

Isolation of bovine blood monocytes

Blood was collected from five cattle (9-month-old charolais-cross steers) by means of venupuncture with tubes containing EDTA as an anticoagulant. Blood was transferred to 50-ml polypropylene tubes and centrifuged at $1400g$ for 20 min at 20 °C. White blood cells were isolated from the buffy coat and mixed with Ca^{2+}- and Mg^{2+}-free phosphate-buffered saline (PBSA) to a final volume of 35 ml. The cell suspension was layered onto 15 ml of 54% isotonic PERCOLL (Amersham Biosciences, GE Healthcare, Baie d'Urfe, Canada) and centrifuged at $2000g$ for 20 min at 20 °C. Peripheral blood mononuclear cells (PBMCs) from the PERCOLL-PBSA interface were collected and washed three times with cold PBSA. Monocytes were purified from isolated PBMCs by purification with $CD14^+$ microbeads (Miltenyi Biotec, Auburn, CA). Monocytes (>95% pure) were plated at 5×10^6 cells/well in six-well plates in RPMI 1640 medium (GIBCO, Burlington, Canada) supplemented with 10% fetal bovine serum (GIBCO). Cells were rested overnight at 37 °C before stimulation with recombinant bovine IFN-γ (100 ng/ml), CpG ODN 2007 (25 μg/ml), or LPS (100 ng/ml).

Kinome array experiments and data collection

Cell pellets were lysed with 80 μl of lysis buffer [20 mM tris-HCl (pH 7.5), 150 mM NaCl, 1 mM EDTA, 1 mM ethylene glycol tetraacetic acid (EGTA), 1% Triton, 2.5 mM sodium pyrophosphate, 1 mM Na_3VO_4, 1 mM NaF, leupeptin (1 μg/ml), aprotinin (1 g/ml), 1 mM phenylmethylsulphonyl fluoride (PMSF)], incubated on ice for 10 min, and then centrifuged in a microcentrifuge at maximum speed for 10 min at 4 °C. An 80-μl aliquot of this supernatant was mixed with 10 μl of the activation mix [50% glycerol, 500 μM ATP, 60 mM $MgCl_2$, 0.05% v/v Brij-35, and bovine serum albumin (BSA, 0.25 mg/ml)] and incubated on the chip for 2 hours at 37 °C in a humidity chamber. After incubation, slides were washed once in PBS-Triton and then submerged in stain (PRO-Q, Diamond Phosphoprotein Stain, Invitrogen, Burlington, Canada) with agitation for 1 hour. Arrays were then washed three times in tubes containing destain [20% acetonitrile (EMD Biosciences, Billerica, MA) and 50 mM sodium acetate (Sigma, Oakville, Canada) at pH 4.0] for 10 min each, with the addition of new destain each time. A final wash was performed with distilled water. A total of 300 peptides and controls were determined as described previously [Jalal et al., 2009]. All peptides were synthesized and printed according to the protocol by JPT Peptide Technologies (`http://www.jpt.com`). Amino-oxyacetylated peptides were synthesized on cellulose membranes in a parallel manner by means of SPOT synthesis technology. Side chains were deprotected, and peptides were cleaved from the cellulose membrane. Peptide solutions were deposited per spot on aldehyde-functionalized glass slides. Peptides were spotted in triplicate on each array. Arrays were dried and read with a GENEPIX professional 4200A microarray scanner (MDS Analytical Technologies) at 532 to 560 nm with a 580 nm filter to detect dye fluorescence. Images were collected and signal collected with GENEPIX 6.0 software (MDS).

7.6.2 Compared methodologies

Much of the kinome analysis published to date adopts methodologies from nucleotide (gene expression) microarray data analysis. To evaluate our proposed methodology, three previously published workflows for microarray analyses were implemented in R and applied to the same data sets and the results compared. Those comparison methodologies are "percentile normalization (PNorm) + fold-change (FC)" [Löwenberg et al., 2006, van Baal et al., 2006, Arsenault et al., 2009, Jalal et al., 2009], "quantile normalization (QNorm) + *limma*" [Schrage et al., 2009], and "VSN + *limma*" [Smyth, 2004]. All three methods operate on background-corrected data. The PNorm procedure was implemented in R based on the algorithm reviewed by Fundel et al. [2008a]. The 90th percentile was used as in the kinome analysis by Löwenberg et al. [2006]. Briefly, after background correction, intensities in each array were divided by the 90th percentile of the data points from the same array so as to achieve a uniform intensity at the 90th percentile across all the arrays. The QNorm and VSN steps were performed with the *limma* function *NormalizeBetweenArrays* by setting the *parameter* method to *quantile* and *vsn*, respectively [Smyth and Speed, 2003]. *NormalizeBetweenArrays* provides the VSN method. After VSN, however, it further scales the transformed data by taking the logarithm to base

2 ($\log_2$), which is not performed in our pipeline. In the "PNorm + FC" approach, data for any peptide with a SD larger than 1.96 times the mean of its replicate data points from the same array were deemed inconsistent and excluded from subsequent analysis [Löwenberg et al., 2006, van Baal et al., 2006]). In "QNorm + *limma*" and "VSN + *limma*," a function called *duplicateCorrelation* in the *limma* package was used to estimate the correlation between the duplicates within an array [Smyth et al., 2005]. The resulting correlations for each peptide were used as a weighting factor for the subsequent differential analysis. Finally, an F-test provided by the *limma* package was used to compare the $\log_2$ fold-changes (logFC) of each peptide across biological replicates. The use of *duplicateCorrelation* and an F-test are not mentioned in the two corresponding studies [Fletcher et al., 2009, Schrage et al., 2009], but both seem to be reasonable steps and should only put the comparison methodologies in a stronger position. In the differential analysis, the "PNorm + FC" approach identifies differentially phosphorylated peptides by comparing their combined FCs with an arbitrary threshold, termed "td." The peptides with FCs larger than +td are deemed to exhibit statistically significant phosphorylation, whereas those with FCs less than -td are classed as being statistically significantly dephosphorylated. The two other comparison methods involving *limma* use the function eBayes [Schrage et al., 2009] to determine P-values associated with moderated t-statistics. A peptide is determined as differentially phosphorylated if its P-value is less than 0.1. For the comparison methodologies, pathway identification was also performed with InnateDB [Lynn et al., 2008]. All of the peptides, except those determined to have inconsistent intensities, were considered. Thresholds were the same as for our new method (P-value of 0.1 and FC value of 1). For "QNorm + *limma*" and "VSN + *limma*," identifiers of the peptides together with P-values and synthetic fold-change values were again input. The log ratios provided by *limma* were converted to FC values with the R function *logratio2foldchange* from the *gtools* library. For "PNorm + FC," only peptide identifiers and FC values were input because no P-values are available from this method. Visualizations of differential phosphorylation patterns were not presented in the studies describing the comparison methodologies, and none was added as part of this work. However, hierarchical clustering and PCA are established techniques that are easily applied to the normalized and filtered intensities from the compared methodologies.

7.6.3 Comparison criteria

The P-values for the overrepresented JAK-STAT, IL-2, and TLR pathways from InnateDB were used as the central criteria for the comparisons between our proposed pipeline and the published methods described earlier. Because of the fairly small total number (300) of different kinase substrates included in our datasets, lenient P-value and FC thresholds for filtering differentially phosphorylated peptides were chosen to increase the chance of discovering meaningful pathways with each of the four methods. Those thresholds were chosen to be 0.1 and 1, respectively.

7.6.4 Data sets

We compiled data sets for three different experiments conducted at different times. The kinome microarray had the same design in all three experiments. Because 300 peptides were spotted on the array and there were three intraarray replicates, 900 signal intensities were obtained from ArrayVision for each experimental run. In the first experiment, monocytes from three outbred cattle labeled "89," "136," and "149" were treated with IFN-γ or media control (denoted as "IFN" and "MonoIFN," respectively, in the subsequent discussion). The second and third experiments examined the kinomic responses of monocytes induced by CpG-containing oligonucleotides and LPS (denoted "CpG" and "LPS," respectively) relative to their individual media controls (denoted "MonoCpG" and "MonoLPS," respectively). Only one treatment replicate was obtained for each of these experiments, with a different animal being used than those indicated for the experiment with IFN-γ.

7.6.5 Data processing before analysis

The raw data exhibited noticeable variance-versus-mean dependence for signals elicited by the 900 peptide spots. This problem occurs when the variances of signal intensities for individual peptides are not constant but increase as mean intensity increases. This can be observed in a graph in which ranks of the 900 means of the peptide signals are plotted against the corresponding SDs (Figure D.1, top left). The dependence was diagnosed as an increasing curve (rising to the right). The systematic trend largely diminished after normalization by any of the four techniques, in addition to a fifth technique of $\log_2$ alone, which was made possible after eliminating negative values that resulted from background correction. Among these methods, the VSN transformations yielded the best results, as indicated by almost horizontal lines (Figure D.1, bottom middle and bottom right). However, the $\log_2$-scaled VSN appears to achieve the best result of the two. We determined a frequency distribution for the data from each normalization (Figure D.2). As is evident, only the transformed data from the $\log_2$-scaled VSN or standalone VSN approached a normal distribution. Distributions derived from other techniques appeared skewed. However, as exemplified by scatter plots of the signal intensities for cells treated with CpG as compared with control cells (Figure D.3), patterns within the responses of the same peptides under any two different treatments in the raw data were better preserved by PNorm and VSN without log-scaling. The patterns were poorly preserved by the $\log_2$-based VSN and the remaining methods (Figure D.3). In conclusion, stand-alone VSN, the transformation that we used in our methodology, appeared to be the method of choice as a preprocessing step.

7.6.6 Spot-spot variability analysis to determine inconsistent peptides

According to step 4 of the PIIKA Methodology section (Supplementary Materials), we performed a χ^2-test on the transformed data. In general, fewer than 11, but more than 1, peptides were inconsistently phosphorylated on a chip (that is, $P < 0.01$ based on the χ^2-test statistic TS1 in each replicate). These peptides were dealt with as described in the Supplementary Materials. In contrast, the comparison method "PNorm + FC"

produced a larger range of inconsistent peptides (from 2 to 28). These inconsistent peptides were manually eliminated from subsequent differential analysis. Although no explicit test for consistency of spot intensities across multiple chips was performed in "QNorm + *limma*" and "VSN + *limma*," the correlations of the technical replicates calculated by *duplicateCorrelation* had subsequent effects on the P-values determined in the corresponding differential analyses.

7.6.7 Subject-subject variability analysis to exclude biological variation

In an outbred species, such as cattle, a degree of individual-to-individual variability in biological responses is observed [Wilkie and Mallard, 1999, Pal and Lewis, 2004]. To identify conserved biological processes, we applied an F-test to the data sets from the IFN-γ experiments from the three animals to determine animal-dependent and animal-independent responses (step 6 of the PIIKA Methodology in the Supplementary Materials). This test was performed after biological subtractions (that is, considering the spot intensities of the cells treated with IFN-γ after subtracting the spot intensities corresponding to the control cells). Under the same treatment condition, any peptide with $P < 0.01$ was considered animal-dependent. By this criterion, four peptides out of 300 (just over 1%) appeared to be animal-dependent and were eliminated in subsequent analysis of IFN data. As a comparison, the F-test from *limma* identified only two animal-dependent peptides (with no overlap with the previous set of four peptides). The proportion from either test seems very low; however, it is a result of the very stringent P-value. Because there were no biological replicates in the experiments in which cells were treated with CpG or LPS, subject-subject variability analysis was not applied to those data sets.

7.6.8 Treatment-treatment variability analysis to calculate the statistical significance of differences in phosphorylation

For all of the methods except "PNorm + FC," we listed the total numbers of differential peptides and the numbers of significantly phosphorylated and dephosphorylated peptides at a 90% statistical confidence (Table 7.2). Because "PNorm + FC" does not calculate a statistical significance for the peptides deemed to be differentially phosphorylated, it was not included in the comparisons. Because of our experimental design, a considerable number of substrates were expected to exhibit statistically significantly different extents of phosphorylation relative to those of the controls. However, both "QNorm + *limma*" and "VSN + *limma*" seem to be overstringent, and they identified only a few kinase targets (Table 7.2). This was especially the case for "VSN + *limma*." In contrast, PIIKA identified a much larger set of differentially phosphorylated peptides under each treatment (Table 7.2). Despite the use of the same data transformation method, the additional logarithmic transformation in the "VSN + *limma*" method led to a statistically significantly different outcome for each treatment.

Table 7.2: The total numbers of differentially phosphorylated peptides at 90% significance level as discovered by three different methods. Differentially phosphorylated peptides from cells treated with CpG, LPS, or IFN-γ were identified by the three methods "QNorm + *limma*," "VSN + *limma*," and PIIKA (our proposed methodology). $\uparrow$ and $\downarrow$ indicate the number of identified peptides with increased or decreased phosphorylation, respectively, with respect to the control condition, and $\updownarrow$ indicates the total number of the two numbers. The "PNorm + FC" method was not included because it does not enable a calculation of the significance of the presence of phosphorylated peptides.

	QNorm + *limma*			VSN + *limma*			VSN + paired t-test (PIIKA)		
Treatments	$\updownarrow$	$\uparrow$	$\downarrow$	$\updownarrow$	$\uparrow$	$\downarrow$	$\updownarrow$	$\uparrow$	$\downarrow$
CpG	11	1	3	3	3	0	85	44	41
LPS	17	11	6	9	5	4	55	28	27
IFN	16	7	9	8	4	4	133	71	62

7.6.9 Visualization of analyzed data

Our proposed methodology includes a visualization scheme for comparative analysis of kinome patterns induced by all three ligands (IFN-γ, CpG, and LPS). The visualization (Figures 7.2 and 7.3), with labels identifying each peptide, depicts the amount of statistical significance of the phosphorylation status of each peptide elicited from bovine monocytes treated by IFN-γ, CpG, or LPS relative to the corresponding controls (the top, bottom left, and bottom right sectors in each circle in Figure 7.3, respectively). The animal-dependent peptides under treatment with IFN-γ identified from the F-test in the analysis of subject-subject variability described earlier are indicated by a gray color in the corresponding upper sectors in the circles on the right in the plot. Peptides with excessive variability across technical replicates for any of the treatments, as determined by the χ^2-test, are represented in white. Statistically significant phosphorylation and dephosphorylation events are presented in yellow and blue, respectively. The color depths are inversely proportional to the corresponding P-values from the one-sided paired t-test. The visualization is laid out in an augmented fashion as described in step 9 of the PIIKA Methodology (Supplementary Materials).

From the plot, it is evident that 74 peptides have common differential phosphorylation status across the three treatments (Figure 7.3, circles from 85 on the top left to 160). Forty-one peptides appear to undergo similar phosphorylation under treatment with CpG and LPS, but not IFN-γ (circles from 3 on the top to 290). These commonly active peptides may be involved in shared signaling pathways specifically induced by CpG and LPS, which both activate receptors of the same family. A higher degree of conservation between the signaling of CpG and LPS, rather than between the signaling of CpG and IFN-γ, would be anticipated. Our group [Arsenault et al., 2009] as well as others [Yi et al., 2001] have demonstrated the initiation of overlapping cellular responses at the levels of phosphorylation-mediated signaling as well as gene expression after activation of immune cells with these ligands. This conservation between CpG and LPS was visually apparent earlier (Figure 7.2), in which about half of the circles have the same dominant color for both

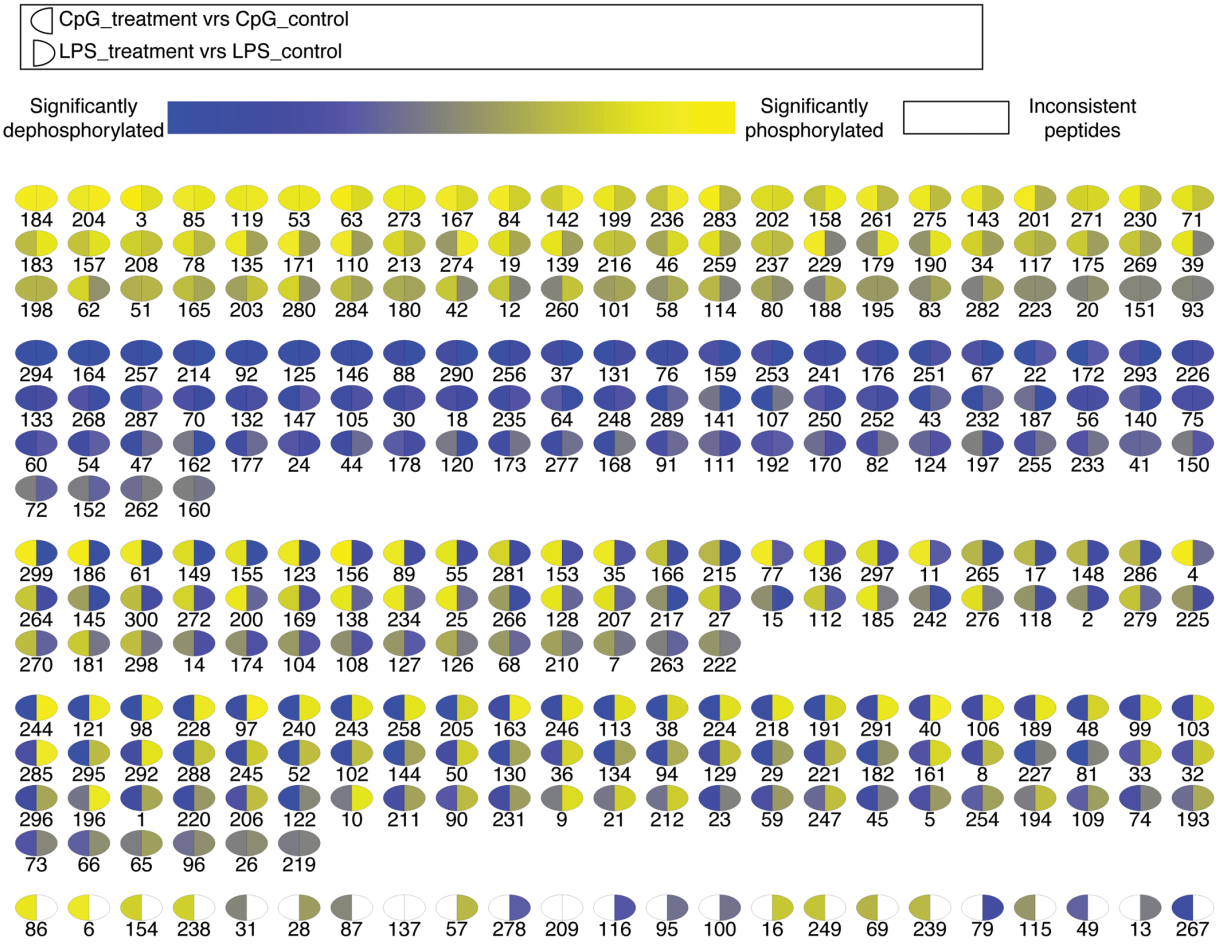


Figure 7.2: Visualization of differential phosphorylation in the CpG and LPS data sets based on the P-values from the one-sided, paired t-test. Each peptide is represented by a colored circle, where the coloration of the left and right semicircles indicates the P-values from the tests of CpG versus MonoCpG (control) and LPS versus MonoLPS (control), respectively. The extents of yellow and blue coloration are proportional to the amount of statistical significance of phosphorylation and dephosphorylation, respectively. White indicates statistically significant spot-spot variability across technical replicates or animal dependency as determined by the χ^2-test and F-test, respectively. The number below each circle is the original position number of the peptide in the microarray. The circles are arranged in blocks, top to bottom, according to whether the peptide is mutually phosphorylated or dephosphorylated in both treatments of the pair, or phosphorylated in one and dephosphorylated in the other. The last block contains peptides inconsistently phosphorylated for at least one treatment of the pair. See step 9 in the PIIKA Methodology (Supplementary Materials) for further information. The coloration has been changed from red and green as described there to blue and yellow to improve clarity and to aid color-blind readers.

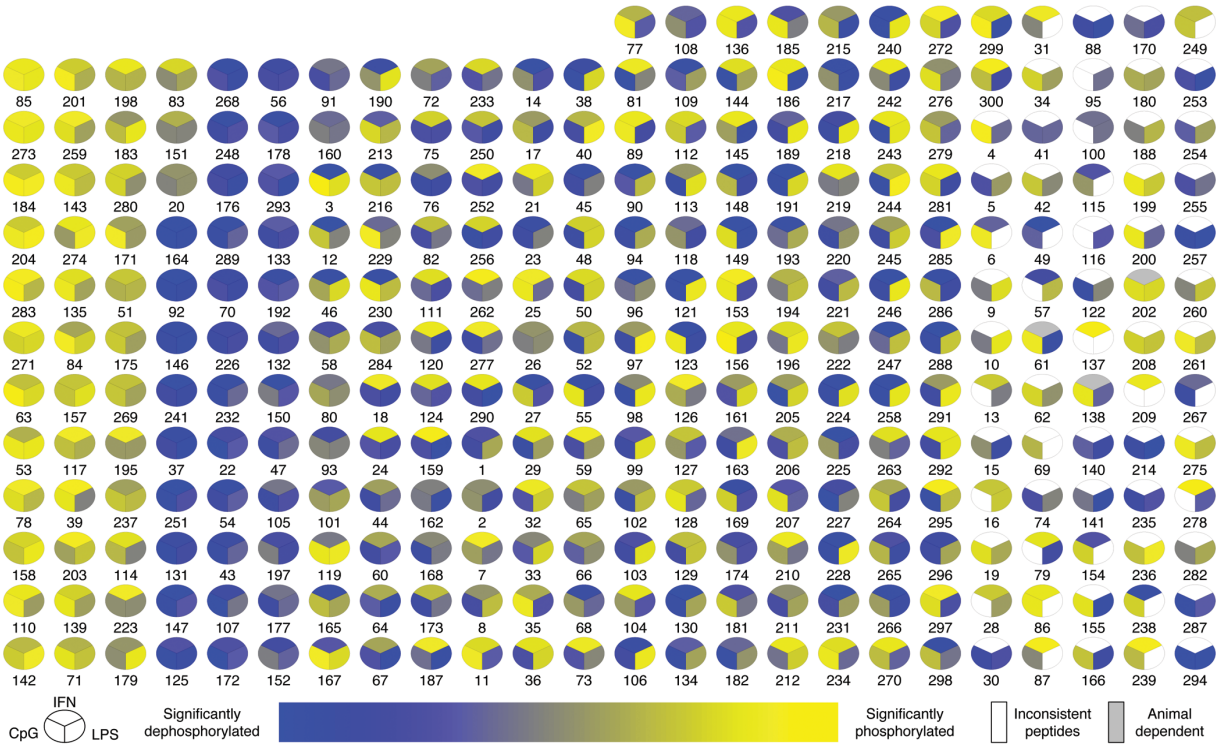


Figure 7.3: Visualization of differential phosphorylation in the three data sets based on the P-values from the one-sided, paired t-test. Each peptide is represented by a colored circle. In each circle, the coloration of top, left, and right sectors indicates the P-value from the tests of IFN versus MonoIFN, CpG versus MonoCpG, and LPS versus MonoLPS, respectively. The extents of yellow and blue coloration are proportional to the amount of statistical significance of phosphorylation and dephosphorylation events, respectively. White indicates spot-spot inconsistency and grey indicates animal dependency as determined by the χ^2-test or F-test, respectively. The number below each circle is the original position number of the peptide in the microarray. The shape of the visualization mimics the physical appearance of the array. The circles are arranged in blocks, left to right, according to whether the peptide is mutually phosphorylated or dephosphorylated in all treatments or phosphorylated in one or some treatment but dephosphorylated in another or other treatments. The last block contains peptides inconsistently phosphorylated for at least one treatment of the trio. See step 9 in the PIIKA Methodology (Supplementary Materials) for further information. The coloration has been changed from red/green as described there to blue/yellow to better accommodate readers with color vision deficiencies.

Table 7.3: Pathway analysis results from InnateDB (`http://www.innatedb.ca`), a publicly available pathway analysis tool. Based on the extents of differential phosphorylation, InnateDB predicts pathways that are consistent with the experimental data. Each pathway is assigned a probability value (P) based on the number of proteins (corresponding to input peptides) present from that pathway. Output includes the number of uploaded peptides associated with a particular pathway as well as the subset of those peptides that are differentially phosphorylated. "Pep" (↕) indicates the number of peptides on the array that relate to the pathway, whereas ↑ and ↓ show the number of identified peptides of the pathway with increased or decreased phosphorylation, respectively, relative to the control condition.

Ligand	Pathway	Pep	PNorm + FC			QNorm + *limma*			VSN + *limma*			VSN + paired t-test (PIIKA)		
		↕	↑	↓	P	↑	↓	P	↑	↓	P	↑	↓	P
CpG	TLR	34	15	12	0.021	4	0	0.019	1	0	1.000	14	4	0.008
LPS	IL-2	25	8	10	0.587	1	2	0.600	1	2	0.405	9	0	0.002
IFN	JAK-STAT	25	18	2	0.122	1	1	0.700	1	0	1.000	12	0	0.003

semicircles.

7.6.10 Identifying signaling transduction pathways with InnateDB

The previous step identified sets of peptides that were differentially phosphorylated under specific conditions. As described in step 11 of the PIIKA Methodology (Supplementary Materials), this information can be used to identify known signaling pathways by using databases such as InnateDB (`http://www.innatedb.ca`) [Lynn et al., 2008]. Therefore, we input identifiers of the peptides in the three data sets into the online database together with the P-values and fold-change values. This was performed with the analysis data from PIIKA as well as from the three comparison methodologies. In response, the query mechanism at the online database provided a list of pathways and associated P-values for the pathways and identified those of the input peptides that appeared in the output pathways. The model signaling pathways known for the ligands CpG, LPS, and IFN-γ are TLR, IL-2, and JAK-STAT, respectively. The numbers of peptides corresponding to proteins in each data set that were found in these model pathways were determined (Table 7.3), as well as the statistical significance of the pathway as calculated from the whole data set by InnateDB. Results indicated an improved level of statistical significance for our analysis pipeline as opposed to the alternative methods, without resulting in an appreciable loss in sensitivity. Specifically, comparing against the best P-value from any of the other three methods, the PIIKA method improves the P-values for the known pathways TLR, IL-2, and JAK-STAT from 0.019 to 0.008, 0.405 to 0.002, and 0.122 to 0.003, respectively. These improved results may be a result of the increased numbers of peptides that were deemed to be differentially phosphorylated by our method (Table 7.3). Visual representations of the respective signaling pathways indicate how our analysis method (the right panel in each row) identifies more proteins in the signaling pathways, creating a more robust network as compared with that of "QNorm + *limma*" (Figure 7.4). We present only the "QNorm + *limma*" method, because it was more accurate and discriminating than was the "VSN + *limma*" method according to our results (Table 7.3) and because no P-value was associated with peptides in "PNorm + FC."

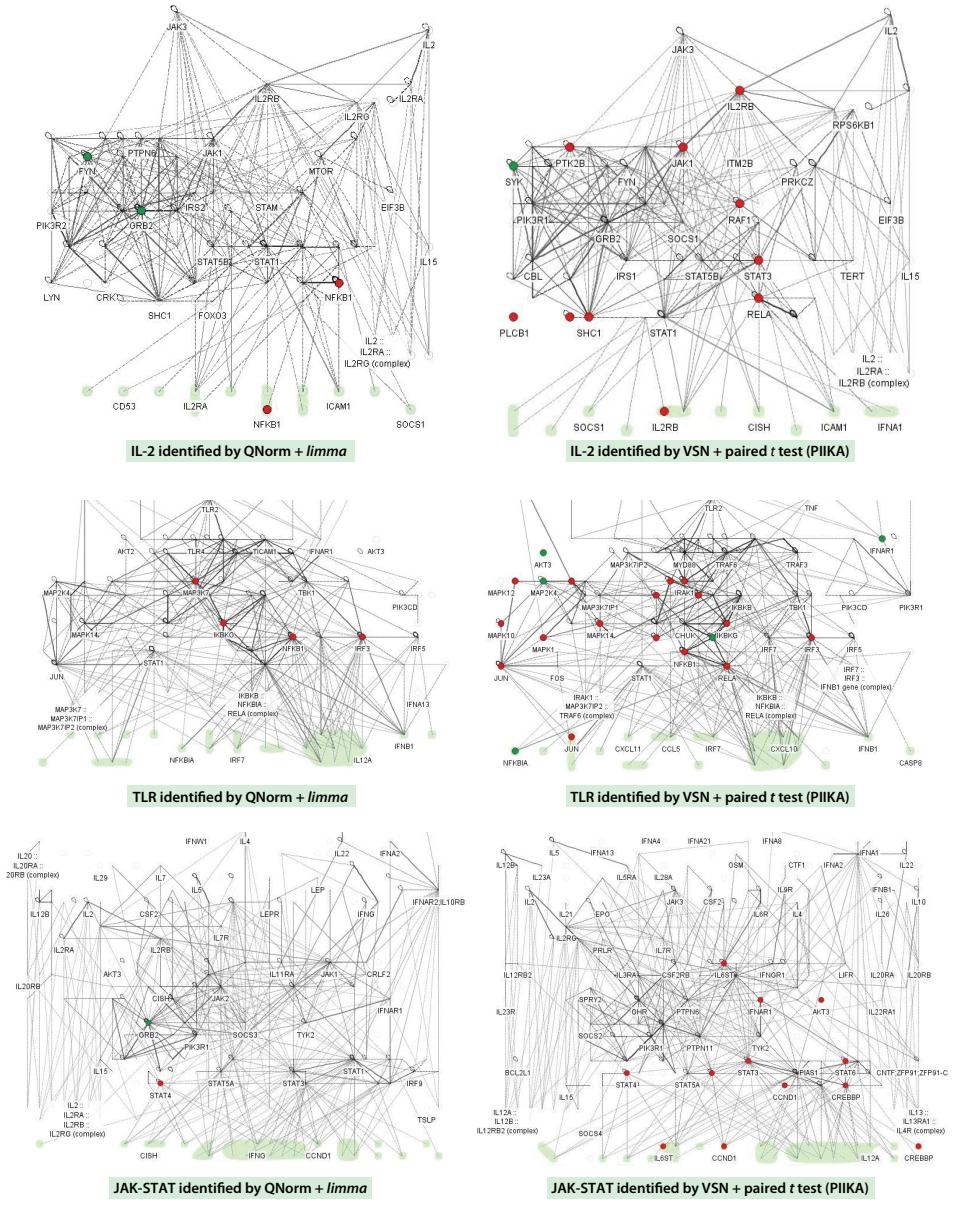


Figure 7.4: Network representations of identified signaling pathways. The top panel shows the IL-2 signaling pathways identified from data from experiments in which cells were treated with LPS. The middle and bottom panels present, respectively, the TLR signaling pathways identified from experiments in which cells were treated with CpG oligonucleotides and the JAK-STAT pathway identified from experiments in which cells were treated with IFN-γ. The nodes in each network represent proteins containing peptides that were identified as statistically significantly differentially phosphorylated. Red coloration of a node indicates an increase in phosphorylation, whereas green indicates a decrease in the extent of phosphorylation. The hue intensity represents the magnitude of the increase or decrease in phosphorylation status. The noncolored spots are either not identified (that is, they were on the array but were not determined to be significantly phosphorylated) or they were not on the array. The networks were generated through the use of the Cerebral plugin [Barsky et al., 2007] for the interaction viewer Cytoscape [Shannon et al., 2003]. The network on the left was derived from "QNorm + *limma*," whereas the network on the right was from PIIKA. The greater abundance of colored (red or green) nodes in the networks on the right indicates that the results from PIIKA led to fuller identification of the relevant pathways as compared to using the results from "QNorm + *limma*."

7.6.11 Clustering analysis of analyzed data to determine treatment-related patterns

The goal of clustering is to make patterns inherent in the kinome data visually evident, which can help uncover biological results and confirm hypotheses suggested by other steps in the methodology. As an example, we performed PCA on the intensity values in the three data sets after VSN transformation and spot-spot and subject-subject variability analysis, according to step 10 of the PIIKA Methodology (Supplementary Materials). One would expect the three pathways represented in the data to exhibit spatial separation in a plot with the three principal components (PCs). This was indeed the case. In the 3D PCA plot, the data sets were widely dispersed along the axis of the first PC (PC1). However, it was also expected that the commonalities in the TLR and IL-2 pathways described earlier would be evident in the PCA plot. This expectation was met, as shown in a 2D plot with the axes PC2 and PC3 (Figure D.4), in which the data points for the CpG and LPS treatments are clustered close together.

7.7 Notes and remarks

Given the similarity in data acquisition techniques between kinome arrays and gene expression nucleotide arrays, it is understandable that data analysis methods previously developed for gene expression data are being used for kinome data. Numerous software packages exist for the interpretation of gene expression data, and many researchers assume that these techniques are also generally applicable to kinome data [Löwenberg et al., 2006, van Baal et al., 2006, Schrage et al., 2009]. However, for the reasons described earlier, the distinct biological nature of kinome data motivates questioning the use of the same systematic approaches as are used for gene expression analysis. For example, the *limma* package is specifically designed to analyze transcription or cDNA arrays. Based on our results (Tables 7.2 and 7.3), it appears that the package may be too conservative in analyzing kinome arrays of only moderate size (for example, 900 spots per chip used in the current data sets). To our knowledge, there have been no systematic studies to determine which gene expression techniques are applicable to kinome data or how they should be modified to deal with kinome data.

We have established a framework to address the challenges presented by kinome microarray data analysis. We selected a set of transformations, statistical tests, and statistical thresholds to address the variability between technical and biological replicates and to identify true differential phosphorylation of a peptide in response to a specific treatment. We then implemented a conforming kinome analysis software pipeline called PIIKA. To comparatively analyze PIIKA, we performed kinome analysis of monocytes stimulated with three different ligands of well-understood signaling pathways. Each data set was analyzed by our methodology and three popular alternative strategies. The results of this comparative analysis suggest that our framework and pipeline offer improved extraction of biologically relevant information in terms of the confidence (P-value) with which signaling pathways are identified as well as the number of phosphorylation events implicating

those pathways.

The signal intensities come from fluorescent dyes that specifically bind to phosphorylated peptides. For peptides that are unphosphorylated or weakly phosphorylated, nonspecific binding of the dye to regions surrounding the peptide may result in background intensities that are higher than those of the foreground. This leads to negative intensity values after background correction [Jalal et al., 2009]. Such negative values were observed in the input data sets (see the example in step 1 of the PIIKA Methodology in the Supplementary Materials). The commonly used workflow from gene expression studies with percentile or quantile normalization, averaging, and fold-change calculations in the differential analysis is not directly applicable to the negative values but nonetheless has been applied to kinome analyses in many studies [Hestvik et al., 2003, Löwenberg et al., 2006, van Baal et al., 2006]. The technique excludes negative values and is therefore subject to information loss. In contrast, our proposed technique uses the VSN transformation that brings all of the data points (including the negative ones) onto the same positive scale while maintaining the correlations between them (Figures D.1, D.2, and D.3, bottom right) [Huber et al., 2002]. Therefore, all of the information from the kinome experiments is preserved by the transformation. Despite starting with the same VSN transformation, the function *NormalizeBetweenArrays* from *limma* applies a further $\log_2$ function over the transformed intensities, which tends to disturb the intrinsic data structure (Figure D.3, bottom middle).

Fundel et al. pointed out that different normalization procedures may have profound effects on the distribution, as well as the statistical significance values, of the extent of gene expression [Fundel et al., 2008a]. This phenomenon also carries over to kinome data. Indeed, the outcomes from QNorm, PNorm, VSN (log-scaled), and VSN alone differed greatly from each other (Figures D.1, D.2, and D.3). Only PNorm and VSN appeared to preserve the inherent correlations between the treated and controlled responses (Figure D.3). Moreover, despite using the same *limma* method, "QNorm + *limma*" and "VSN + *limma*" yielded substantially different outcomes in differential analysis (Table 7.2), further demonstrating the importance of the choice of transformation procedure.

Witten and Tibshirani have noted that in the analysis of microarray data, there is no correct answer as to whether fold-change or the modified t-statistic should be used [Witten and Tibshirani, 2007]; however, the choice can have a dramatic effect on the set of genes or peptides that are identified. Therefore, the measure used must be based on the biological system under investigation. Specifically, if large absolute changes are relevant to the system, then fold-change should be used; on the other hand, if changes relative to the underlying noise are important, then P-values or modified t-statistics (for example, the paired t-test in our pipeline) are preferable. On the basis of the outcomes of the comparison of the four methodologies here, it appears that the appropriate transformation and statistical tests selected for PIIKA enabled the pipeline to elucidate biologically meaningful signaling pathways in all three treatments.

7.7.1 Future work

In the analysis of microarray data, a single data set is used multiple times to accept or reject hypotheses. For example, in our methodology many one-sided paired t-tests were performed on the basis of the same set of preprocessed signal intensities. This is an instance of the statistical problem of multiple hypothesis testing. To complicate matters, the paired t-tests assume that the extents of phosphorylation of the peptides are independent. However, independence is not guaranteed because several peptides with different phosphorylation sites may come from the same protein; thus, phosphorylation of these sites may be correlated. To deal with this multiple testing situation, techniques such as Bonferroni or Benjamini can be used [Benjamini and Hochberg, 1995, Montgomery, 2009]. The Bonferroni technique, for example, is applicable to both independent and dependent tests but is susceptible to producing more false negatives when the tests are dependent. A potential problem with these techniques is an overstringency that is imposed to achieve a small type I statistical error (for example, 5%). This is typically not a problem for the analysis of gene expression data, in which tens of thousands of genes are considered at one time and an aim of the analysis is to reduce dimensionality. In that case, high specificity is favored over sensitivity; false positives are avoided at the cost of more false negatives. However, the dimensionality of kinome data sets is smaller than that of transcriptome data sets, and phosphorylation of peptides may not be as efficient as the hybridization of oligonucleotides on transcription arrays *in vitro* [Jalal et al., 2009]. Therefore, it is advisable to less readily eliminate peptides because some of them may turn out to be crucial in the pathway analysis, as has been exemplified in our results (Tables 7.2 and 7.3). In general, dealing with the multiple-hypothesis testing problem in the context of kinome microarray data warrants further investigation.

An optional argument to the *vsn2* function in step 2 of the PIIKA Methodology (Supplementary Materials) is a model to be used as the basis for the transformation. This parameter is not supplied in the current version of PIIKA; thus, the VSN transformation uses the entire input data set as its model. An alternative model can be specified to, for example, normalize data to a specific reference set so that data from multiple experiments can be combined. Exploration of this option, and identification of appropriate reference sets, is also part of future work.

In addition to the implementation of PIIKA as an R script, a prototype has been developed as a Web-based server and corresponding graphical user interface (GUI). The interface is written in PHP and Javascript and requires only a standard Web browser. This interface is being ported to run within a Galaxy [Goecks et al., 2010] user environment, and that version will be made available in the future.

7.8 Funding

This work was supported by a Canadian Institutes of Health Research (CIHR) Catalyst grant (to Scott Napper), a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant (to Anthony Kusalik), NSERC graduate scholarships (to Brett Trost and Ryan Arsenault), and NSERC research

assistantships (to Yue Li and Jillian Slind). Philip Griebel is a holder of a Tier I Canada Research Chair in Mucosal Immunology, which is funded by CIHR.

Chapter 8

PIIKA 2: An expanded, web-based platform for analysis of kinome microarray data

Brett Trost, Jason Kindrachuk, Pekka Määttänen, Scott Napper, and Anthony Kusalik

This is the second of two papers that relate to the analysis of kinome microarray data. It describes a new version of PIIKA, PIIKA 2, which contains many new features, primarily in the areas of statistical analysis, clustering, and data visualization.

Citation

B. Trost, J. Kindrachuk, P. Määttänen, S. Napper, and A. Kusalik. PIIKA 2: an expanded, web-based platform for analysis of kinome microarray data. *PLOS ONE* 8(11):e80837, 2013.

Author contributions

All authors helped conceive and design the experiments, participated in analyzing the data, and helped revise the paper. Brett Trost performed the experiments. Scott Napper contributed reagents and materials. Brett Trost wrote the majority of the paper, with additional contributions by Jason Kindrachuk.

Supplementary material

This paper is accompanied by five supplementary files numbered S1 through S5. Supplementary File S1, which is a PDF file, is reproduced in Appendix E. Supplementary Files S2 through S5 are large tables or text files, and can be accessed via `http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0080837`.

8.1 Abstract

Kinome microarrays are comprised of peptides that act as phosphorylation targets for protein kinases. This platform is growing in popularity due to its ability to measure phosphorylation-mediated cellular signaling in a high-throughput manner. While software for analyzing data from DNA microarrays has also been used for kinome arrays, differences between the two technologies and associated biologies previously led us to develop Platform for Intelligent, Integrated Kinome Analysis (PIIKA), a software tool customized for the analysis of data from kinome arrays. Here, we report the development of PIIKA 2, a significantly improved version with new features and improvements in the areas of clustering, statistical analysis, and data visualization. Among other additions to the original PIIKA, PIIKA 2 now allows the user to: evaluate statistically how well groups of samples cluster together; identify sets of peptides that have consistent phosphorylation patterns among groups of samples; perform hierarchical clustering analysis with bootstrapping; view false negative probabilities and positive and negative predictive values for t-tests between pairs of samples; easily assess experimental reproducibility; and visualize the data using volcano plots, scatterplots, and interactive three-dimensional principal component analyses. Also new in PIIKA 2 is a web-based interface, which allows users unfamiliar with command-line tools to easily provide input and download the results. Collectively, the additions and improvements described here enhance both the breadth and depth of analyses available, simplify the user interface, and make the software an even more valuable tool for the analysis of kinome microarray data.

8.2 Introduction

Catalyzed by protein kinases, reversible protein phosphorylation is the most widespread signaling mechanism in eukaryotes and plays a critical role in virtually every cellular process. Technologies for studying phosphorylation-mediated signaling in a high-throughput manner have the potential to facilitate the discovery of complex biomarkers, help identify signaling pathways associated with particular diseases, and provide general information regarding regulatory mechanisms. One such technology is the kinome microarray, in which natural substrates of protein kinases are mimicked by short (15-mer) peptides containing the phosphoacceptor site (at the central position) as well as the same surrounding residues as in the corresponding intact protein. The phosphorylation kinetics of these peptides and their corresponding proteins are similar [Zetterqvist et al., 1976, Kemp et al., 1977]. First proposed in 2002 [Houseman and Mrksich, 2002, Houseman et al., 2002], kinome arrays have since been used to study a large variety of biological systems, such as the effects of glucocorticoids on the immune system [Löwenberg et al., 2005], signaling in chondrosarcoma [Schrage et al., 2009], sugar signaling in plants [Ritsema et al., 2009, Ritsema and Peppelenbosch, 2009], stem cell differentiation [Hazen et al., 2011], bacterial infections in cows [Arsenault et al., 2013a, Määttänen et al., 2013], and many others [Peppelenbosch, 2012].

Previously, researchers using kinome microarrays have analyzed the resulting data using software designed for DNA microarrays. However, the chemistry involved in the two technologies is different, and data processing appropriate for one technology may not be appropriate for the other. Further, given the smaller number of spots on a kinome array (∼300-1000) versus a DNA array (∼30,000), the use of the same statistical stringency thresholds commonly employed in DNA array software could compromise the ability to identify differentially phosphorylated peptides in kinome arrays and to identify changes in the modulation of biological pathways. DNA microarray software also often lacks statistical techniques for ascertaining the consistency of technical and biological replicates. In response to these concerns, we developed a software program in the R environment [R Development Core Team, 2006] called Platform for Intelligent, Integrated Kinome Analysis (PIIKA) [Li et al., 2012], and showed that it improves the ability to identify cellular signaling pathways that are upregulated or downregulated in response to a particular treatment. PIIKA also facilitates the identification of peptides that have inconsistent responses among the technical replicates on a single array or among different biological replicates (e.g., different animals exposed to the same treatment), ensuring that only high-quality data are used in subsequent statistical and clustering analyses.

Here, we report the development and release of PIIKA 2, which contains many additions and improvements to PIIKA, primarily in the categories of cluster analysis, statistical analysis, and data visualization. Among others, PIIKA 2 allows users to perform the following tasks, which would have been impossible in the original PIIKA without substantial user effort (e.g., writing of scripts).

- determine the statistical significance of the consistency between the actual clustering of the data and a hypothesized clustering;
- identify subsets of peptides that induce a particular clustering;
- assess the statistical significance of hierarchical clustering nodes using bootstrapping analysis;
- quickly access false negative rates and positive and negative predictive values for the t-tests between pairs of samples;
- easily evaluate the technical and biological reproducibility of the experiment;
- visualize principal component analysis (PCA) results using a three-dimensional interactive plot;
- visualize points that are both statistically significant and have high fold-change values using volcano plots; and
- view the relationships between the normalized signal intensities in pairs of samples.

In summary, PIIKA 2 improves the ability to answer complex biological questions about kinome array data and to make informed decisions concerning statistical thresholds and significance. Whereas the original PIIKA was available only as a command-line tool, PIIKA 2 may also be used via a web-based interface,

which eases the data analysis process for users unfamiliar with the use of command line tools. A significant advantage of PIIKA 2 over stand-alone graphical user interface (GUI)-based tools is that there is no need to click on menu items and change options for each individual analysis the user would like to perform. PIIKA 2 performs all analyses that are applicable given the input provided by the user and outputs the results in the form of spreadsheet-compatible text files and publication-ready images.

As mentioned, PIIKA 2 is available in two forms: a web-based version, and a local version that can be installed on the user's computer. Both versions are available through the Saskatchewan PHosphorylation Internet REsource (SAPHIRE) website at `http://saphire.usask.ca`. PIIKA 2 is free for academic use; users interested in PIIKA 2 for commercial purposes should contact the authors.

The remainder of this paper is divided into three major sections. The Methods section discusses the methodology associated with each new feature of PIIKA 2. The Results section gives examples and figures that illustrate the application of these features to data from a real kinome microarray experiment. Finally, the Discussion and conclusion section summarizes the value of PIIKA 2 for analyzing kinome array data and discusses the utility of kinome arrays for signaling research in general.

8.3 Methods

When dealing with complex data such as those arising from kinome microarrays, asking non-trivial questions of the data often requires expertise in mathematics, programming and data visualization—as well as a significant investment of time. Ultimately, these often deter users from interrogating their data to the full extent possible. To address this problem, we have implemented in PIIKA 2 a rich assortment of analysis tools. These tools relate to cluster analysis, statistical analysis, or data visualization. As we receive feedback from users, other functionality will be added. This section contains descriptions of the methodologies used; for examples of the use of these methodologies, including relevant figures and example outputs, see the Results section.

8.3.1 Cluster analysis

The original version of PIIKA allowed users to perform hierarchical clustering on the samples in a given experiment; however, the tools available to analyze the clusters were limited. Here, three features new to PIIKA 2 are described that allow users to perform more detailed analyses of their hierarchical clustering results.

Random tree analysis: statistical significance of the clustering of *a priori* groups

In many kinome microarray experiments, the samples or treatments can be placed *a priori* in different groups based on either biological knowledge or specific attributes of the samples or treatments. For brevity, in the following discussion the members of these groups will be called "samples", although if each experimental

treatment has more than one sample associated with it, then the members of these groups would more accurately be called "treatments".

In a real experiment conducted by our research group, for example, one sample was taken from each of 6 biological subjects at each of 4 time points. These samples were then processed using kinome microarrays containing 297 unique peptides, each replicated 9 times on the same array. Image analysis software was used to capture the phosphorylation intensity of each spot as described previously [Jalal et al., 2009], and the resulting data were processed using PIIKA 2. The exact nature of the experiment, the samples, and the subjects is not relevant here (a manuscript describing these data from a biological perspective is in preparation); in this study, the critical feature of the example experiment is that we hypothesize that samples from the same subject will have similar kinome profiles. The original version of PIIKA included functionality for performing hierarchical clustering, which allows the similarity of the kinome profiles of the samples to be ascertained. Although one can get a sense of whether the expected clustering pattern does indeed exist by visually inspecting the resulting dendrogram, this does not give a measure of statistical significance. To remedy this, PIIKA 2 allows the question, "Do samples from the same group cluster together better than would be expected by chance?" to be addressed by deriving an empirical statistical distribution and then reporting a P-value based on this distribution, where a small P-value indicates that samples within the same group (in the above example, the same biological subject) cluster together better than would be expected at random.

Since each step in the process of performing hierarchical clustering results in a bifurcation, clusterings made in this way can always be represented as binary trees. For ease of reference, we therefore convert the dendrogram representation to its corresponding binary tree representation. To evaluate the "goodness" of clustering for a given binary tree T, we define a metric $\delta(T)$ wherein larger values denote better clustering. Suppose that, in our hypothesized grouping of the samples, there are n groups labeled $G_1, G_2, \ldots, G_n$, each containing m samples. In the example above, $n = 6$ and $m = 4$. Also, let the internal nodes of T be labeled $I_1, I_2, \ldots, I_k$, where k is the number of internal nodes. We define a function $f(i, j)$ as follows:

$$f(i,j) = \begin{cases} 0 & \text{if } I_i \text{ has any leaves as descendants that correspond to a group other than } G_j \\ w & \text{otherwise, where } w \text{ is the number of descendant leaves of } I_i \text{ corresponding to group } G_j \end{cases}$$

Then

$$\delta(T) = \sum_{j=1}^{n} \max_{1 \le i \le k} f(i,j) \tag{8.1}$$

In other words, to calculate $\delta(T)$, for each group G_j we find the internal node I_i with the greatest number of leaves as descendants that correspond to G_j and that has no leaves corresponding to any other group. The number of such leaves is added to $\delta(T)$. Thus, the maximum possible value of $\delta(T)$ is nm, and the possible values of $\delta(T)$ are the integers between 0 and nm. To make the metric independent of n and m, it can be

expressed as a ratio: $\delta'(T) = \frac{\delta(T)}{nm} \times 100$. A $\delta'(T)$ value of 100 indicates perfect clustering. The Results section contains an example of a tree T and the calculation of its corresponding score $\delta'(T)$.

While $\delta'(T)$ by itself gives a sense of the goodness of clustering, it does not indicate whether the samples from each *a priori* group cluster together better than would be expected at random. To determine this, 10,000 random trees $R_1, R_2, \ldots, R_{10000}$ are generated (the number of random trees generated can be changed by the user), and the value of δ' is calculated for each. The random trees are generated by modifying the original data matrix, wherein rows represent peptides and columns represent arrays, by randomly rearranging the values within each column. The values $\delta'(R_1), \delta'(R_2), \ldots, \delta'(R_{10000})$ represent an empirical probability distribution for δ'. Thus, the P-value is simply the proportion of random trees R_i for which $\delta'(R_i) \geq \delta'(T)$. For each R_i, PIIKA 2 outputs the rearranged matrix that was used to produce that random tree, visual and text-based representations of the hierarchical clustering of that matrix, and the value of $\delta'(R_i)$. PIIKA 2 also outputs $\delta'(T)$ and the aforementioned P-value.

Peptide subset analysis: identifying sets of peptides that support the clustering of *a priori* groups

Given a set of groups of samples defined *a priori* based on biological knowledge or other factors, it may also be of interest to identify sets of peptides for which the phosphorylation patterns are similar within samples from the same group and different between samples from different groups (as described above, the members of the groups may be either samples or treatments, but for brevity we will just call them "samples"). In other words, one might want to identify sets of peptides for which the clustering of the samples into these groups is as close to perfect as possible. For example, consider a hypothetical experiment in which cell extracts are taken from mice with a genetic propensity to a certain disease, and that we divide these mice into two groups—those that eventually get the disease, and those that do not. If we could identify a set of, say, 10 peptides that have similar responses in mice of the same group, and different responses between groups, then these 10 peptides could potentially act as a biomarker for this disease.

PIIKA 2 implements this functionality using a simple local search procedure. First, the samples (or treatments, if more than one sample corresponds to a particular treatment) are hierarchically clustered using a set of exactly two peptides drawn from the complete set. The score for the corresponding tree (which, again, is a clustering of the samples, not the peptides), $\delta'(T)$, is then determined. This procedure is then repeated for all possible pairs of peptides. The pair of peptides which results in the tree with the greatest value of $\delta'(T)$ is then selected as the "seed". If more than one set has the same value of $\delta'(T)$, then one of them is arbitrarily chosen to be the seed. A third peptide is then added to this list by scanning the remaining peptides and determining which one—in addition to the two chosen as the seed—results in the set with the greatest value of $\delta'(T)$. Additional peptides are iteratively added in the same fashion until all peptides have ultimately been added, in which case the dendrogram is identical to the one created using all of the peptides. For each iteration, the hierarchical clustering is performed anew (as opposed to adding the next peptide onto

the structure of the previous tree).

PIIKA 2 outputs, for each i ($3 \leq i \leq p$, where p is the number of peptides), the dendrogram containing i peptides, the score $\delta'(T)$ associated with that dendrogram, and a spreadsheet-compatible table showing the names of those peptides as well as their normalized intensity values for each sample. The peptides forming these subsets are those having phosphorylation patterns that are similar within samples from the same group, but different between samples from different groups. Depending on the biological application, it might be of interest to examine small sets of peptides (say, 5 or 10) that have this property, or it might be more meaningful to examine larger sets of peptides. The output of PIIKA 2 allows the user to examine sets of peptides with any cardinality between 3 and the total number of unique peptides.

Bootstrap analysis of hierarchical clustering

When performing hierarchical clustering, the strength of the support for each cluster can be ascertained using bootstrapping. As a complement to the heatmaps produced by PIIKA, PIIKA 2 also outputs dendrograms showing the hierarchical clustering of the samples, with each node labeled with two P-values: the bootstrap confidence P-value (BP) as proposed by Felenstein [1985], and the approximately unbiased P-value (AU) as proposed by Shimodaira [2002, 2004]. Each P-value ranges between 0 and 100, and represents the percentage of times that the cluster appears in the bootstrap replicates. The R package `pvclust` [Suzuki and Shimodaira, 2006] is used to calculate these bootstrap values and generate the graphical version of the dendrogram.

It should be noted that the variables (peptides) are not strictly independent, largely because a given kinase might catalyze the phosphorylation of several peptides on the array. This could compromise the statistical soundness of the bootstrap analysis, as each resampling of the original data may not reflect the dependence originally present among the variables. However, similar bootstrap analyses have successfully been used for DNA microarrays [e.g., Finak et al., 2006, Ben-Porath et al., 2008, Ebert et al., 2009, Singh et al., 2009, Ojalvo et al., 2010]), despite the fact that the expression levels of individual genes may not be independent (due, for example, to transcription factors that each promote the transcription of several genes). This suggests that bootstrap analysis should be valuable for kinome arrays as well. Nonetheless, the fact that the peptides are not independent should be kept in mind when interpreting the results.

8.3.2 Statistical analysis

In the original version of PIIKA, several statistical tests were provided, including a t-test for comparing treatment-control combinations, a χ^2-test for identifying peptides inconsistently phosphorylated among the technical replicates, and an F-test for determining the consistency of biological replicates. In this section, we describe statistical analyses performed by PIIKA 2 that were not possible to perform in the original PIIKA.

False positive and false negative probabilities

The original version of PIIKA allowed the user to select a value for α (the probability of a type I error; also called the false positive rate) for the t-tests done between each peptide for a given treatment and control. While controlling the type I error rate is important, it is also important to be cognizant of the type II error rate (denoted β, and also called the false negative rate). This is particularly true because subsequent analyses often involving feeding the data into a program like InnateDB [Lynn et al., 2008], which examines whether a particular cellular signaling pathway appears to be upregulated or downregulated based on the increased or decreased phosphorylation of individual components of that pathway. If the false negative rate is too high, then peptides that are differentially phosphorylated may not be correctly identified, causing pathways to be missed that are in fact differentially regulated in the treatment condition compared to the control condition. As such, it could be valuable to the user to display these false negative probabilities.

In its output files that give the t-test results for each peptide for each treatment-control combination, PIIKA 2 now also includes the value of β for each peptide. These values are calculated using the R package `pwr`. Since β decreases when α is increased, the user can choose to increase the value of α if the values of β are judged to be too high. Note that increasing the number of intra-array technical replicates will also lower the false negative probabilities, although this is usually not an option at the stage in the experiment where array data have already been gathered.

Positive and negative predictive values

Let A represent the event of rejecting the null hypothesis, and let N represent the event that the null hypothesis is true. Then the false positive probability α can be defined as $P(A|N)$. While α is a useful quantity, sometimes it is more meaningful to know the complementary probability $P(N|A)$ (sometimes called "positive predictive value")—given that we rejected the null hypothesis, what is the probability that it is true? $P(N|A)$ can be calculated mathematically using Bayes' rule: $P(N|A) = P(A|N) \times P(N)/P(A)$. Both $P(A|N)$ and $P(A)$ are easy to determine: $P(A|N) \equiv \alpha$, which is supplied by the user, while $P(A)$ is the proportion of peptides attaining a P-value less than α. Unfortunately, $P(N)$ is more difficult to determine, as this represents the actual background probability that a particular peptide will not be differentially phosphorylated. PIIKA 2 uses a (somewhat arbitrary) default value of 0.75 for this value, although this can be changed by the user if desired.

Similarly, it may also be useful to find the probability that the null hypothesis is false given that we failed to reject it (sometimes called "negative predictive value")—that is, $P(\overline{N}|\overline{A})$. Analogous to the above, this can be determined using Bayes rule: $P(\overline{N}|\overline{A}) = P(\overline{A}|\overline{N}) \times P(\overline{N})/P(\overline{A})$. Here, $P(\overline{A}|\overline{N}) \equiv \beta$, while $P(\overline{N})$ and $P(\overline{A})$ are just the complements $P(N)$ and $P(A)$, respectively.

As with β, the t-test files produced by PIIKA 2 now include the probabilities $P(N|A)$ and $P(\overline{N}|\overline{A})$ as described above. $P(\overline{N}|\overline{A})$ is given as a column in the file, as it potentially will differ for each peptide; however, $P(N|A)$ will have the same value for every peptide, so it is listed in a separate file.

Technical and biological reproducibility summaries

To facilitate statistical hypothesis testing, kinome arrays typically contain between three and nine intra-array technical replicates; in other words, between three and nine distinct spots are placed on the array for each unique peptide sequence. In the original PIIKA publication [Li et al., 2012], we described the use of a χ^2-test to identify peptides that are inconsistently phosphorylated among the technical replicates on a single array.

In our own publications describing results from biological experiments involving kinome microarrays [e.g., Määttänen et al., 2013], we typically include a statement summarizing the technical reproducibility of the phosphorylation signal for all the arrays used in the experiment. For instance, for arrays that each contain 300 unique peptides, we might claim that the average number of consistently phosphorylated peptides on a given array was 288, and that this value ranged from 282 to 297. In the previous version of PIIKA, the user would have had to manually calculate these values from other output. However, PIIKA 2 generates a file containing the number of consistently phosphorylated peptides for all the arrays in the experiment, along with the average value and range of values, making it easy to include this information in a manuscript describing the experiment.

In addition to summarizing technical reproducibility, PIIKA 2 also summarizes the biological reproducibility if the experiment involves more than one biological replicate per treatment. The information presented is analogous to that given in the technical reproducibility summary: for each treatment, the number of peptides consistently phosphorylated among the biological replicates is given, along with the average and range of these values.

8.3.3 Data visualization

The original version of PIIKA contained three major data visualization methods: heatmaps (showing the hierarchical clustering of samples on the x axis and peptides on the y axis), 2-dimensional and 3-dimensional scatterplots showing the results of PCA, and a novel visualization method for comparing differential phosphorylation P-values between two treatment-control combinations [Li et al., 2012]. PIIKA 2 provides several additional visualization methods; these are described below.

PCA visualization using Virtual Reality Modeling Language

While the first three principal components can be visualized using a 3D scatterplot, as provided in the original PIIKA, it can be difficult to comprehend such plots, especially when there are many samples. The layout of sample labels can also pose problems in 3D scatterplots. As such, interactive plots created using virtual reality modeling language (VRML) are an attractive alternative. PIIKA 2 uses the R package `vrmlgen` [Glaab et al., 2010]—specifically, the function `cloud3d`—to generate 3D scatterplots in VRML. Using an appropriate viewer, such as Instant Player (`http://www.instantreality.org`), the user can rotate and translate the figure, as well as zoom in and out, making the relationship between the samples in three-dimensional space

much easier to comprehend.

Volcano plots

When comparing the level of phosphorylation between a treatment and a control, two quantities are often of interest: the P-value corresponding to the t-test, which answers the question, "Is there a statistically significant difference between the phosphorylation level in the treatment and the phosphorylation level of the control?", and the fold-change (FC) value, which answers the question, "What is the magnitude of the difference between the phosphorylation level in the treatment compared to the control?". These quantities are not necessarily meaningful in isolation: very large or very small FC values may be associated with a lot of variability in the technical replicates, and thus have an insignificant P-value according to the t-test; conversely, the magnitude of the difference between the treatment and control may be small, but the technical replicates may be highly consistent within each sample, leading to a small P-value. A useful visualization method for looking at both fold-change values and P-values simultaneously is the "volcano plot" [Cui and Churchill, 2003]— a scatterplot with FC on the x-axis and P-value on the y-axis, and named as such because the pattern exhibited by the points usually resembles an erupting volcano. Points located in the upper-left or upper-right corners of the plot are usually of the most interest, as they have both small P-values and high FC values. PIIKA 2 generates a volcano plot for each treatment-control combination specified by the user.

Scatterplots between pairs of samples

In addition to visualizing how different samples are from each other using hierarchical clustering or PCA, it may be useful to compare the normalized intensity values between two samples at a more fine-grained level—i.e., by directly visualizing differences in responses between individual peptides. To facilitate this, PIIKA 2 outputs, for each possible pair of samples, a scatterplot containing a point for each peptide, where a point's x and y coordinates represent that peptide's normalized intensity value for the first and second sample in the pair, respectively. Each scatterplot also contains a least-squares regression line, the line $y = x$ (for comparison to the regression line), and a statement giving the Pearson correlation between the normalized intensity measurements in each sample.

8.3.4 Other features

As a complement to the hierarchical clustering analysis, which may use either Euclidean distance or (1 - Pearson correlation) as the distance metric, PIIKA 2 also outputs files containing the Euclidean distance and Pearson correlation between each pair of samples, as well as each pair of subtracted treatment-control combinations. It may also be of interest to consider the distance between samples or treatment-control combinations by including in the calculation only peptides that are differentially phosphorylated. PIIKA 2 outputs files containing these data as well, with a peptide being considered differentially phosphorylated for

a given pair of treatments or treatment-control combinations if the P-value according to the paired t-test is less than the user-specified threshold.

While PIIKA 2 contains many features related to the analysis and visualization of kinome microarray data, some users may want to perform analyses not available in PIIKA 2 or use their own visualization software. To facilitate this, PIIKA 2 outputs a file for each stage in the analysis pipeline containing the processed data at that stage. Specifically, a file is generated containing the data after background subtraction; after applying the `vsn` transformation; after rearranging the matrix; after averaging the technical and biological replicates; and after performing biological subtraction (if applicable). These files can easily be used as input to external programs.

8.3.5 PIIKA 2 availability

PIIKA 2 is available both as a web server and as a stand-alone program that the user can run on his or her own computer. Each version has the same functionality, and can be accessed or downloaded via the SAskatchewan PHosphorylation Internet REsource (SAPHIRE) webpage at `http://saphire.usask.ca`.

The web-based version of PIIKA 2 is ideal for users who have limited experience with command line-based tools. To use the web-based version of PIIKA 2, the user must upload one or more input files, and enter the value of several parameters (number of intra-array replicates, number of peptides on the array, and so on). Detailed instructions for formatting the input files and choosing parameters are available on the PIIKA 2 webpage. The user must also enter his or her e-mail address; once the job has finished running, the user will receive an e-mail containing a link where the results can be downloaded. A full guide to the output of PIIKA 2 is available in Appendix E; a continuously updated version of the output guide is available via the SAPHIRE website, and will also be included along with the other results files that the user downloads once their job is complete.

Commercial providers of kinome microarrays usually offer custom-designed arrays, where the client chooses the number of unique peptides to include on the array, the number of intra-array technical replicates per unique peptide, and the sequences of those peptides. Some providers also offer off-the-shelf arrays, for which the above attributes are predefined. To ease the submission process for those using the latter type, the PIIKA 2 website contains a drop-down menu where the user can select a particular off-the-shelf array. Once selected, the fields for certain parameters (the number of unique peptides on the array and the number of technical replicates per unique peptide) will be automatically filled in with the appropriate values. To identify off-the-shelf kinome arrays, we searched the websites of major providers of peptide arrays, including JPT Peptide Technologies (`http://www.jpt.com`), Pepscan (`http://www.pepscan.com`), Arrayit (`http://www.arrayit.com`), and PEPperPRINT (`http://www.pepperprint.com`).

The stand-alone version of PIIKA 2 is suitable for users familiar with command line-based tools, and requires that the R language [R Development Core Team, 2006] be installed, as well as several R packages. A full guide to installing and running the stand-alone version of PIIKA 2 is included in the download.

8.4 Results

8.4.1 Cluster analysis

Random tree analysis: statistical significance of the clustering of *a priori* groups

To demonstrate the algorithm described in Methods, we use the aforementioned experimental data consisting of one sample taken at 4 time points from 6 subjects. The kinome array data were processed using the usual PIIKA pipeline (background subtraction followed by normalization and transformation using *vsn* [Huber et al., 2002]). Peptides that were consistently phosphorylated across the technical replicates according to a χ^2-test for all 24 arrays ($n = 165$) were then subjected to hierarchical clustering using (1 - Pearson correlation) as the distance metric and average linkage as the linkage method. The resulting heatmap is shown in Figure 8.1, with the sample (column) dendrogram showing that samples from the same subject tended to cluster together quite well, although not perfectly. The question is, do samples from the same subject cluster together better than would be expected by chance?

In our technique for ascertaining the statistical significance of the clustering of predefined groups, a hierarchical clustering is represented as a binary tree. As an example, the binary tree corresponding to the clustering shown in Figure 8.1 is shown in Figure 8.2. In applying Equation 8.1 to this tree, let subject A be G_1, subject B be G_2, and so on. Then $\max_{1\leq i\leq k} f(i,1) = 3$, where k is the number of internal nodes. This expression is maximized when $i = 10$, because internal node I_{10} contains no leaves as descendants that correspond to any group other than G_1 (subject A), and has three leaves as descendants that do correspond to G_1 (the most of any internal node that satisfies the above condition). Similarly, $\max_{1\leq i\leq k} f(i,2) = 3$, $\max_{1\leq i\leq k} f(i,3) = 3$, $\max_{1\leq i\leq k} f(i,4) = 2$, $\max_{1\leq i\leq k} f(i,5) = 3$, and $\max_{1\leq i\leq k} f(i,6) = 3$. The sum of these is 17, and so $\delta(T) = 17$ and $\delta'(T) = \frac{\delta(T)}{nm} \times 100 = \frac{17}{6\times 4} \times 100 = 70.8$.

To generate the distribution of scores that would result by random chance, 10,000 random trees were generated by randomly rearranging the values for each peptide within a given array (column). The value of $\delta'(T)$ was calculated for each of these random trees, and the distribution of these data is shown in Figure 8.3. The lowest score given to a random tree was 0, while the greatest was 58.3. As such, none of the random trees had a score equal to or greater than the score for the actual tree, giving a P-value of less than 0.0001. This indicates that samples from the same subject do indeed cluster together better than would be expected by chance.

Peptide subset analysis: identifying sets of peptides that support the clustering of *a priori* groups

The local search procedure described in Methods was tested using the same sample data as described above. This procedure was used to identify sets of peptides that, when subjected to hierarchical clustering, resulted in a clustering with a value of $\delta'(T)$ as close to 100 as possible—that is, a clustering where the arrays

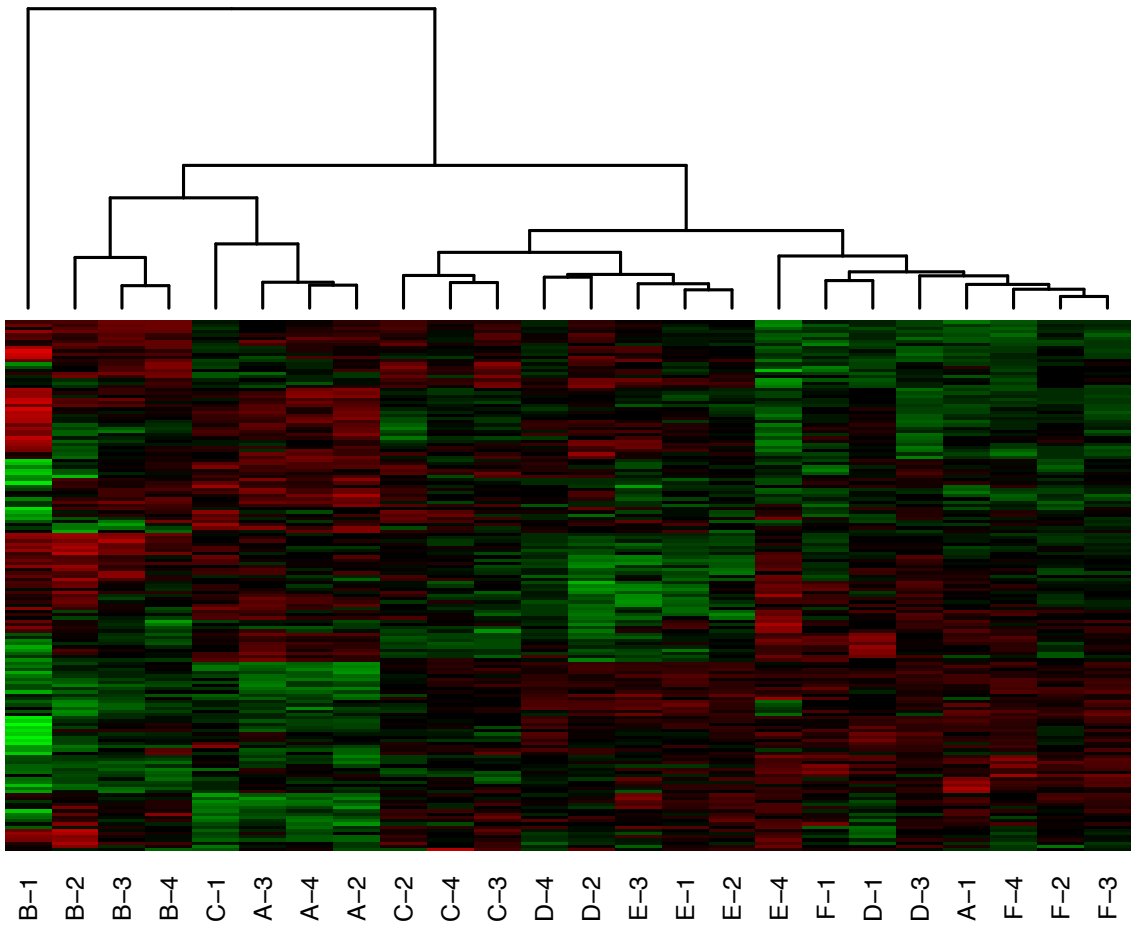


Figure 8.1: Heatmap and hierarchical clustering of kinome microarray profiles from the example experiment. Samples were taken at four time points from six different subjects, here labeled A-F. The number of the sample from the same subject represents the time point at which the sample was taken; for example, sample C-3 was taken from subject C at time point 3. The distance metric used for clustering was (1 - Pearson correlation), while the linkage method used was average linkage.

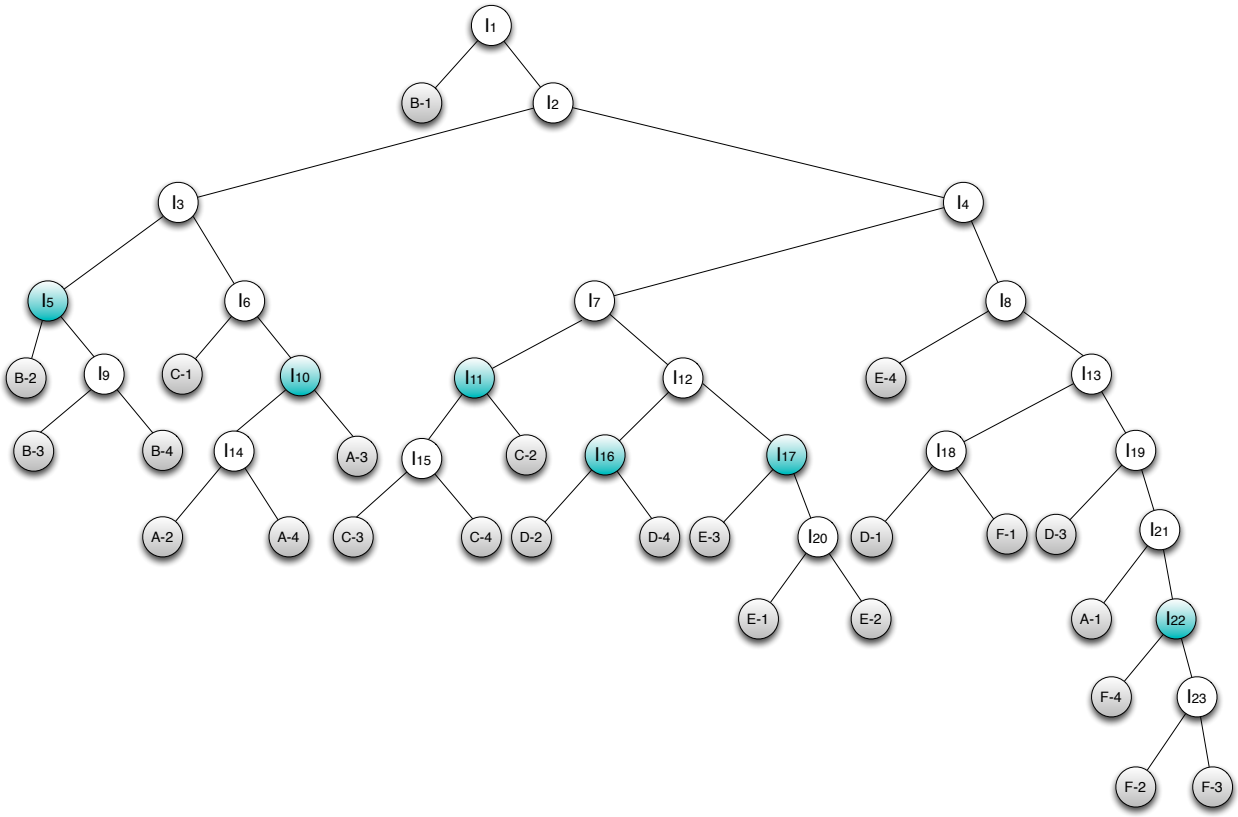


Figure 8.2: Binary tree representation of the dendrogram shown in Figure 8.1. Leaf nodes are shaded in grey and are labeled according to the subject and time point as in Figure 8.1. Internal nodes are labeled I_1 through I_{23}, and those internal nodes I_i for which $f(i, j)$ is maximized for some group G_j (where G_1 corresponds to subject A, G_2 corresponds to subject B, and so on; see also Equation 8.1) are shaded in blue.

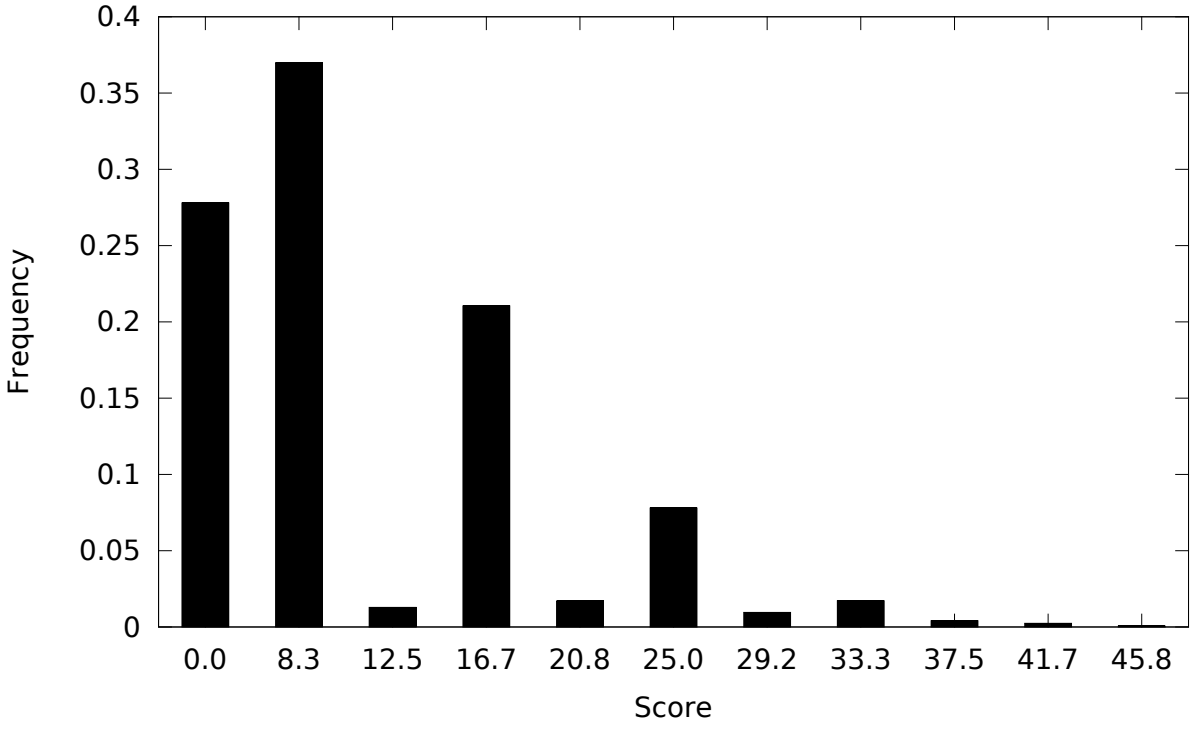


Figure 8.3: Empirical distribution of random tree scores. Ten thousand random matrices $R_1, R_2, \ldots, R_{10000}$ were created from the matrix used to create the sample dendrogram in Figure 8.1 by randomly rearranging the peptide intensity values within each sample. For each score $\delta'(R_i)$ that was given to at least one random tree, the frequency of that score is indicated.

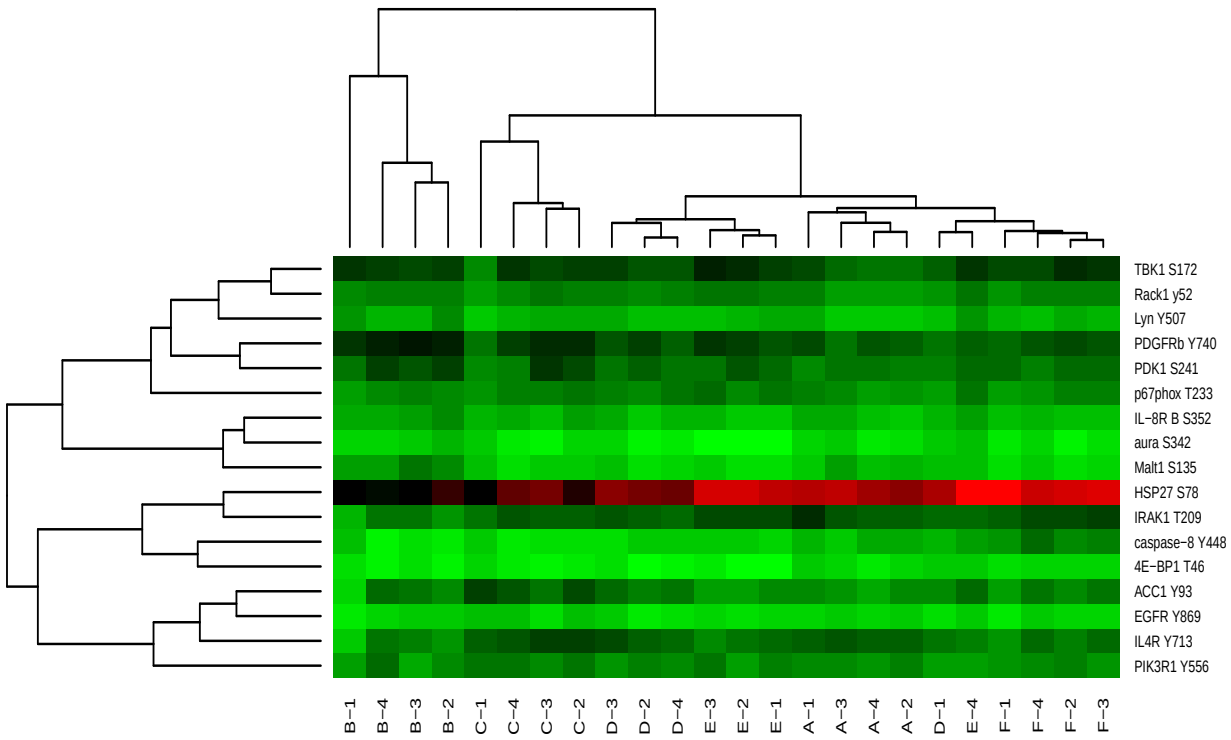


Figure 8.4: Heatmap and hierarchical clustering of kinome microarray profiles of samples from the example experiment using 17 peptides chosen according to a local search algorithm. The same distance metric and linkage method were used as in Figure 8.1. The sample names are the same as in Figure 8.1; the peptide names are also indicated on the right side of each row.

corresponding to a given subject cluster together, and cluster separately from arrays corresponding to other subjects. The greatest score $\delta'(T)$ given to a dendrogram for some number of peptides i was 91.7, which was the case for $11 \leq i \leq 17$. In other words, for each i between 11 and 17 inclusive, a dendrogram could be created with i peptides that had a score of 91.7. The dendrogram corresponding to $i = 17$ is shown in Figure 8.4. Figure 8.4 shows that, as its score suggests, the clustering with these 17 peptides is almost completely concordant with the "ideal" clustering by subject. Specifically, subjects A, B, C, and F all clustered together perfectly, while three out of the four samples from each of subjects D and E clustered together. As such, these 17 peptides were consistently phosphorylated within the same subject, but differentially phosphorylated between subjects.

Bootstrap analysis of hierarchical clustering

One caveat with hierarchical clustering is that clusters are always produced, even in the extreme case where there is no relationship among any of the samples; as such, dendrograms containing bootstrap values represent valuable tools for the user to be able to assess the strength and significance of the clusters produced. PIIKA 2 uses the R package `pvclust` [Suzuki and Shimodaira, 2006] to generate dendrograms with bootstrap P-values

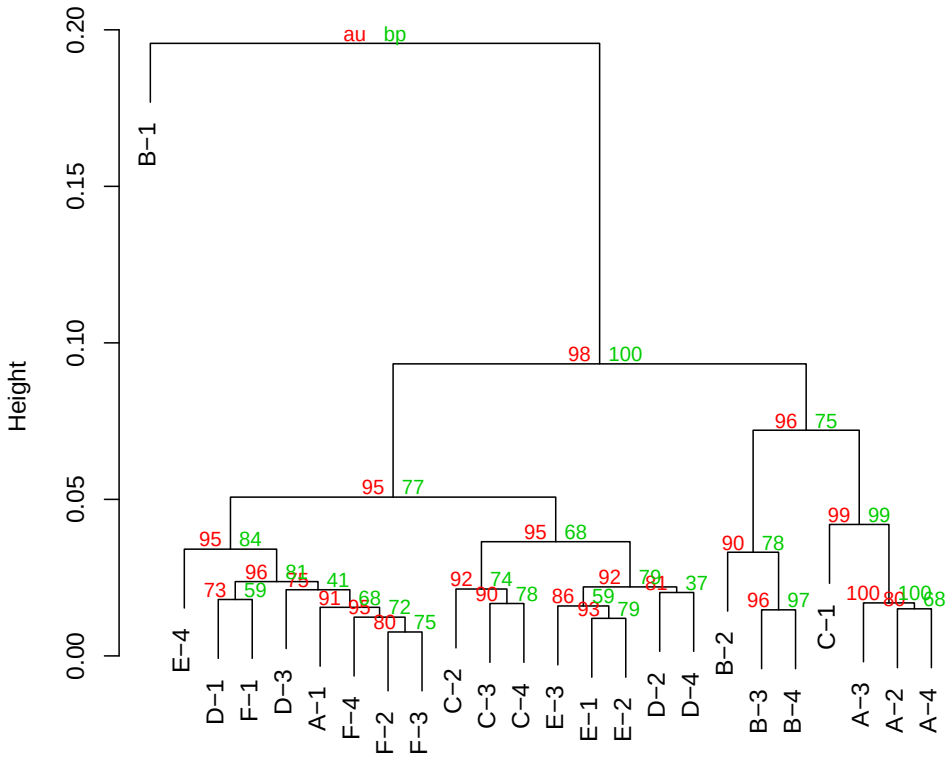


Figure 8.5: Example of a dendrogram with bootstrap values using PIIKA 2. The clustering of the samples is the same as in Figure 8.1. The red numbers represent the approximately unbiased (AU) P-values as determined using the method of Shimodaira [Shimodaira, 2002, 2004], while the green numbers represent the standard bootstrap P-value [Felenstein, 1985]. All calculations and the drawing of the figure were performed using the R package `pvclust` [Suzuki and Shimodaira, 2006].

on each node. These P-values are actually displayed as confidence values on the plot; for instance, a value of 99 means that the null hypothesis ("the cluster is not real") can be rejected at a significance level of 0.01. An example of such a dendrogram, which was created using the same data and clustering methodology as the sample (column) dendrogram in Figure 8.1, is shown in Figure 8.5. For some of the subjects, the samples from the second, third, and fourth time points clustered together, while the sample from the first time point was an outlier (e.g., subject A). Figure 8.5 shows that, for some subjects, we could be very confident in the clustering of the latter three samples. For example, the cluster containing samples from the second, third, and fourth time points for subject A had a confidence value of 100. Conversely, there was somewhat less confidence for subject F, with the cluster containing the same three time points having an approximately unbiased confidence value of 95 but a standard bootstrap value of just 72.

8.4.2 Statistical analysis

False positive and false negative probabilities

As described in Methods, PIIKA 2 now outputs values for β (the false positive rate) for each peptide for each treatment-control combination. These are present in the same files that contain the fold-change and t-test results. An example of such a file is given as Supplementary File S2.

Positive and negative predictive values

In addition to values for β, PIIKA 2 now also outputs positive and negative predictive values—the former being specific to a given treatment-control combination, and the latter being specific to each peptide within a given treatment-control combination. Like β, the negative predictive values are present in the file containing the fold-change and t-test results; see Supplementary File S2 for an example. Since the positive predictive value does not depend on the peptide, a separate file containing just the positive predictive value is generated for each treatment-control combination.

Technical and biological reproducibility summaries

As it is often of interest to determine and summarize the level of reproducibility of the intra-array technical replicates in a kinome microarray experiment, PIIKA 2 outputs a file containing the number of peptides for which the phosphorylation signal was determined to be consistent according to a χ^2-test for each array, as well as the range and average of these values. Supplementary File S3 contains an example of one of these files. If the experiment involves more than one biological replicate per treatment, then the level of reproducibility of these replicates may also be of interest; an example of such a summary given as output by PIIKA 2 can be found in Supplementary File S4.

8.4.3 Data visualization

PCA visualization using Virtual Reality Modeling Language

A (static) picture of a VRML plot generated by PIIKA 2, as rendered by the visualization software Instant Player, is shown in Figure 8.6, and the corresponding VRML file is available as Supplementary File S5. The user has the option of assigning colours to each point in order to categorize them by treatment group, subject, etc. The user can also hover their mouse pointer over a given point to reveal the label corresponding to that point, as well as its coordinates (a three-tuple representing the values corresponding to the first, second, and third principal components, respectively). Collectively, these features should allow users to more easily identify patterns in their data.

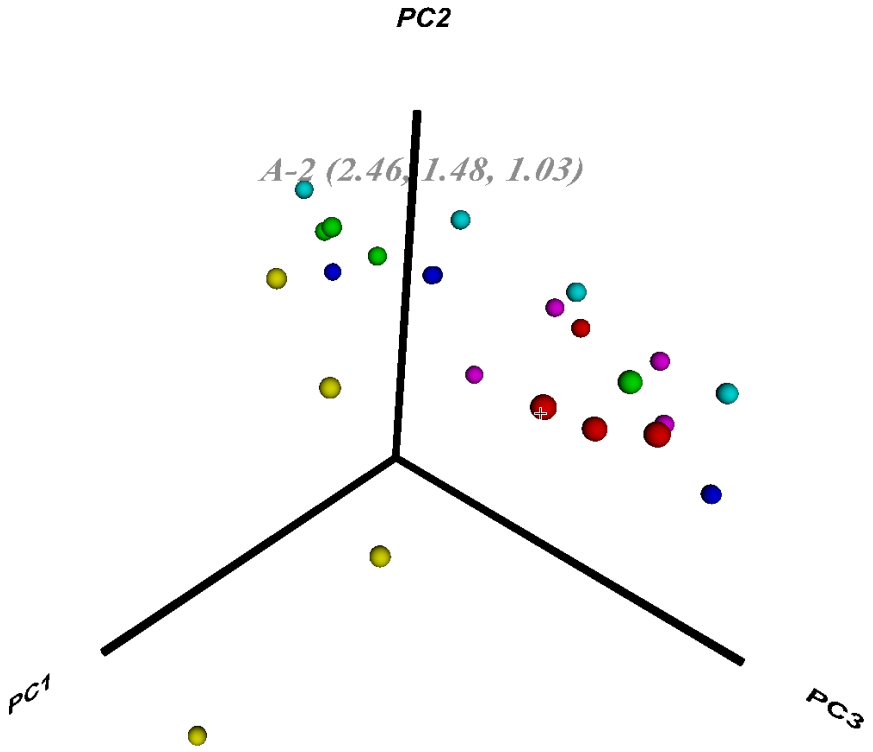


Figure 8.6: Example of a PCA plot generated in VRML format by PIIKA 2. In this experiment, samples were taken from subjects labeled A, B, C, D, E, and F. Samples corresponding to subject A are in red, subject B are in yellow, and so on. The label near the top of the figure is the result of hovering the mouse over the leftmost red circle, and shows that the first, second, and third principal components for this sample had the values 2.46, 1.48, and 1.03, respectively. This image is an example of the visualization given using the VRML viewer Instant Player (`http://www.instantreality.org`).

Table 8.1: Off-the-shelf kinome microarrays that the PIIKA 2 web interface allows the user to select.

Company	Array name	Product code	# technical replicates	# peptides
JPT	Annotated Phosphosites-Kinase	KIN-MA-PhK	9	720
Pepscan	PepChip Kinomics Array	PCKINOM01	3	1024
Pepscan	PepChip Kinase Array	PCKF00020	2	1184
Pepscan	Kinase Evaluation Slide	PCKT00010	2	192

Volcano plots

For a given treatment-control combination, a volcano plot allows the user to easily identify peptides that both have a large FC value and have a significant P-value according to a t-test. An example of a volcano plot generated by PIIKA 2 is given in Figure 8.7. Each point has a specific colour depending on its FC value and P-value (see figure legend). In addition, all points having $|\text{FC}| \geq 2$ are labeled with their respective peptide names, allowing the user to easily identify peptides of interest.

Scatterplots between pairs of samples

Figure 8.8 shows a sample scatterplot produced by PIIKA 2. The red and blue lines represent the diagonal ($y = x$) and the least squares regression line, respectively. The Pearson correlation coefficient is also shown below the x-axis label.

8.4.4 PIIKA 2 availability

PIIKA 2 is available as a web server and as a stand-alone version, both of which can be accessed via `http://saphire.usask.ca`. Figure 8.9 contains a screenshot of the web server. As described in Methods, the web interface includes an option for the user to select an off-the-shelf kinome array purchased from a commercial provider, which allows the fields for certain parameters to be filled in automatically. Of the commercial providers mentioned in Methods, only JPT and Pepscan appeared to offer off-the-shelf kinome arrays, with JPT offering one array appropriate for use with PIIKA 2 and Pepscan offering three. Details on these arrays are given in Table 8.1. This feature will be expanded as more off-the-shelf commercial arrays become available.

8.5 Discussion and conclusion

Many cellular processes can be regulated independently of changes in transcription or translation through post-translational modifications, the most important of which is kinase-mediated protein phosphorylation. Protein kinases play critical roles in regulating complex systems, underlie various pathologies, and represent high-priority drug targets; as such, there is considerable interest in defining and characterizing their biological

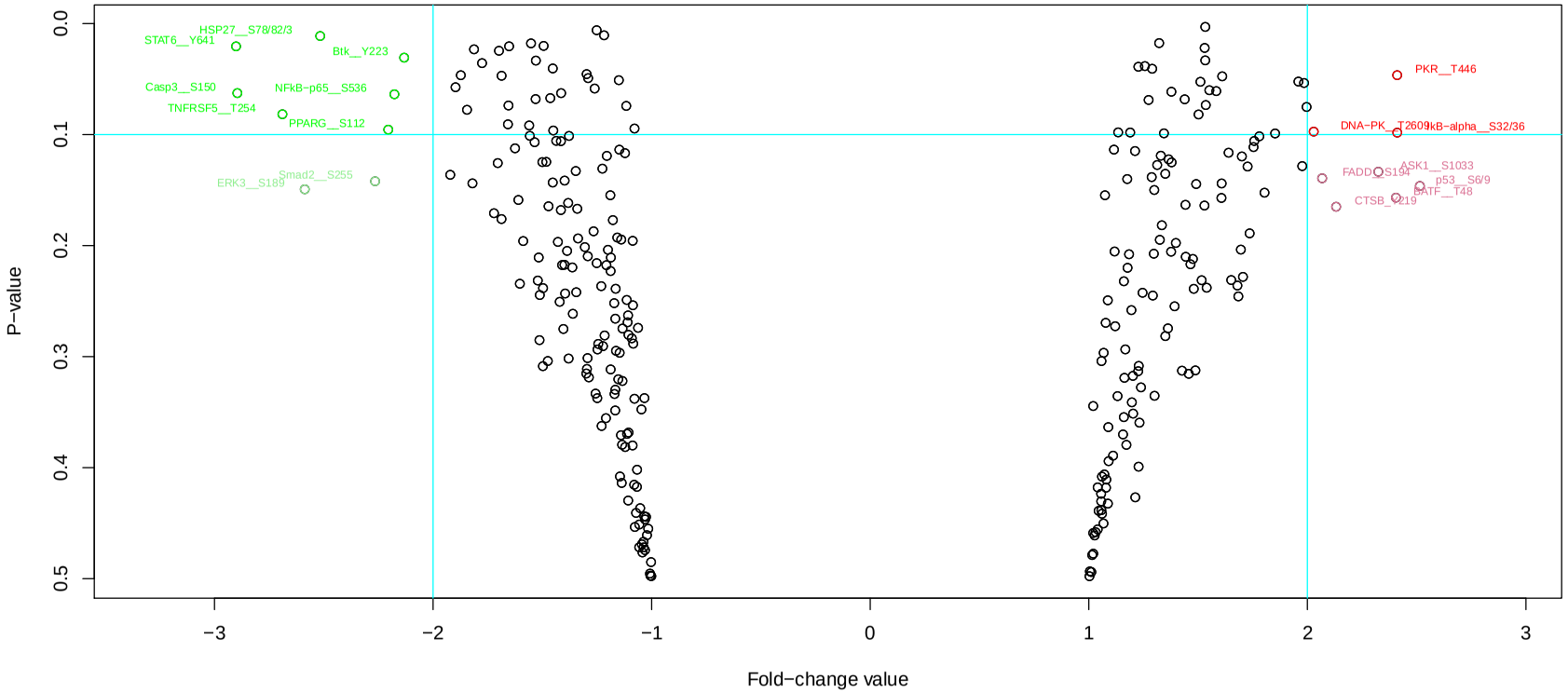


Figure 8.7: Example of a volcano plot generated using PIIKA 2. The points represent actual FC and t-test P-values between two samples from a kinome microarray experiment. Points for which FC ≥ 2 and P-value ≤ 0.1 are coloured red, while those with FC ≥ 2 but P-value < 0.1 are pale red; similarly, points with FC ≤ -2 and P-value ≤ 0.1 are green, while those with FC ≤ 2 but P-value < 0.1 are pale green. All other points are coloured black. The horizontal and vertical blue lines represent the P-value and FC cutoffs, respectively. All coloured points are accompanied by labels showing to which peptide the point corresponds.

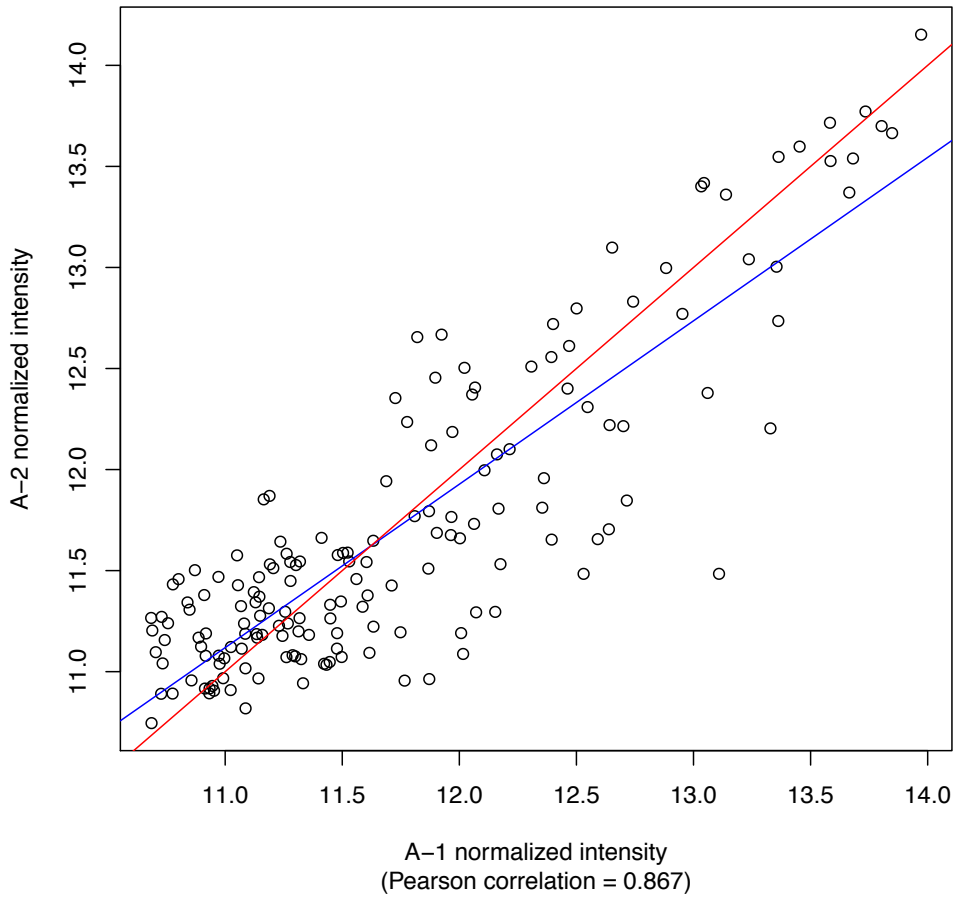


Figure 8.8: Example of a sample-sample scatterplot generated using PIIKA 2. Each point represents a peptide, and the x and y values of that point represent the normalized intensity values for that peptide for the first sample (A-1) and the second sample (A-2). The blue line represents the best fit using least squares, whereas the red line simply shows the diagonal ($y = x$). The Pearson correlation between the two samples is also indicated.

Want to run PIIKA 2 on your own computer instead of using the web-based version? Click here to download the stand-alone version.
Click here for help regarding the files and parameters listed below.
The sample files mentioned below correspond to the sample data mentioned in the paper describing PIIKA 2.

Step #1: Input files

Choose File no file selected — **(Required) Main input file**
Contains the intensity values for your arrays.
Sample file

Choose File no file selected — **(Optional) Treatment-control combinations**
Specifies the treatment-control combinations in your dataset.
Sample file

Choose File no file selected — **(Optional) Treatment-control combinations for P-value visualizations**
Specifies the treatment-control combinations for P-value visualization files.
Sample file

Step #2: Required parameters

Number of technical replicates per unique peptide on the same array
Number of treatments
Number of unique peptides on the array
Number of inter-array replicates
Note: if you entered a value greater than 1 for this option, please use the button below to indicate whether the inter-array replicates represent biological replicates or technical replicates.
Biological

Step #3: Optional parameters

1 - Pearson correlation — Distance metric for hierarchical clustering
McQuitty linkage — Linkage method for hierarchical clustering
Yes — Perform chi-square test?
No — Perform F-test?
Applies only if your dataset contains 2 or more biological replicates.
No — Perform biological subtraction before performing F-test?
Applies only if you are performing an F test.
No — Perform random tree analysis?
Note: if you selected Yes for this option, please enter the number of random trees to generate (default = 10000).
10000
No — Perform peptide subset analysis?
0.1 — Value of alpha (false positive rate) for statistical significance testing
0.25 — Estimated background probability that a peptide will be differentially phosphorylated

Step #4: E-mail address

(Required) Please enter your e-mail address here. Once your job is finished running, you will receive an e-mail with a link where you can download the results. Please note that your e-mail address may be saved for the purposes of tracking usage and of informing you of updates and bug fixes to PIIKA 2.

Step #5: Submit!

Submit

Image credit: Flickr user wildxplorer.

Figure 8.9: Screenshot of the user interface of the PIIKA 2 web server.

roles. Kinome analysis offers three key advantages over traditional profiling of gene and/or protein expression: 1) individual kinase activities are often reliable indicators of phenotypic changes, 2) kinase profiling offers insight into cellular responses at the level of signaling networks, and 3) as kinases are highly "druggable", increased understanding of their biological roles could aid therapeutic design and development.

The growing interest in kinases in both basic and translational research has driven efforts to develop technologies that facilitate the characterization of phosphorylation-mediated signal transduction. Peptide arrays are a relatively inexpensive technology that can be applied to study phosphorylation-mediated cellular signaling in a high-throughput manner. We and other groups have previously demonstrated the utility of kinome arrays for addressing a wide range of biological problems [e.g., Löwenberg et al., 2005, Jalal et al., 2009, Schrage et al., 2009, Ritsema et al., 2009, Ritsema and Peppelenbosch, 2009, Hazen et al., 2011, Kindrachuk et al., 2012, Arsenault et al., 2012]. Given the substantial volume of data generated by kinome arrays, the ability to employ them effectively requires the existence of appropriate analysis methods. In this paper, we have described PIIKA 2, which is a powerful suite of tools for analyzing kinome microarray data. The new analysis tools have significant breadth, covering cluster analysis, statistical analysis, and data visualization. Further, we have provided an online submission platform that allows researchers to easily use PIIKA 2 for their own kinome investigations.

In this paper, the new features in PIIKA 2 were illustrated using a dataset derived from the application of kinome microarrays to real biological samples. However, few details about these samples were given, as this paper focuses on illustrating the capabilities of PIIKA 2, rather than reporting biological conclusions stemming from the analysis of this dataset. However, it should be emphasized that the value of PIIKA 2 lies primarily in its ability to help provide insight into biological systems. A concrete example of this is a recent study by our group that examined the kinome profiles of calf intestinal segments that were either infected or not infected with the bacterium *Mycobacterium avium* subsp. *paratuberculosis* [Määttänen et al., 2013]. In this study, PIIKA 2 was used to show that a given calf's kinome responses clustered into one of two groups, and the specific group to which a given calf belonged correlated with whether the animal exhibited primarily an antibody immune response or primarily a cell-mediated immune response.

As with any software package, future work will relate to the improvement or expansion of existing features, as well as the addition of new features. Several of the additions and improvements to PIIKA 2 were inspired by, or have been useful for, our own research involving the application of kinome microarrays to biological problems. However, some of the questions other researchers wish to address may be different from our own. As such, we are interested in hearing from users of PIIKA 2 regarding ideas for additional features, as well as ways to improve the software in general.

8.6 Supporting information

Supplementary File S1—A guide to the output of PIIKA 2, listing all of the files produced by PIIKA 2, how they are organized, and what information is contained in each file.

Supplementary File S2—A sample file containing results of a statistical comparison (fold-change values, P-values resulting from a paired t-test, values of β, etc.) between a pair of samples from the example experiment.

Supplementary File S3—A sample file containing a summary of the technical reproducibility of the arrays in the example experiment.

Supplementary File S4—A sample file containing a summary of the reproducibility of the biological replicates in the example experiment.

Supplementary File S5—A file in VRML format containing a 3D scatterplot of the first three principal components resulting from principal component analysis. This file can be viewed using any VRML viewer, such as Instant Player (`http://www.instantreality.org`).

8.7 Acknowledgments

We would like to thank Arnie Berg for assistance with R coding, Qingxiang Yan for reviewing parts of the manuscript, and Stephen Johnson and Erin Scruten for helping test the software. This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Division of Intramural Research, NIAID, NIH.

Chapter 9

Kinotypes: stable species- and individual-specific profiles of cellular kinase activity

Brett Trost, Jason Kindrachuk, Erin Scruten, Philip Griebel, Anthony Kusalik, and Scott Napper

This is the first of three papers that describe biological applications of the work described in this thesis. It shows that when blood samples from humans and pigs were subjected to kinome microarray analysis, samples from the same species clustered together far more closely than would be expected at random. This establishes the existence of a species-specific profile of protein kinase activity, or "kinotype". Additionally, within a species, different samples from the same individual clustered more closely than would be expected at random, establishing the existence of individual-specific kinotypes. These observations may have applications in the use of model organisms and in personalized medicine.

Citation

B. Trost, J. Kindrachuk, E. Scruten, P. Griebel, A. Kusalik, and S. Napper. Kinotypes: stable species- and individual-specific profiles of cellular kinase activity. *BMC Genomics* 14:854, 2013.

Author contributions

Scott Napper and Philip Griebel conceived the study. Erin Scruten performed the peptide array experiments. Brett Trost performed the majority of the data analysis, with additional contributions by Jason Kindrachuk, Anthony Kusalik, and Scott Napper. Brett Trost, Jason Kindrachuk, Anthony Kusalik, Philip Griebel, and Scott Napper interpreted the data. Brett Trost, Jason Kindrachuk, and Scott Napper wrote the paper, with substantial revisions by Anthony Kusalik. All authors read and approved the final manuscript.

Supplementary material

This paper is accompanied by five supplementary files. Each of these is a large text or spreadsheet file, and can be accessed via `http://www.biomedcentral.com/1471-2164/14/854/additional`.

9.1 Abstract

Background: Recently, questions have been raised regarding the ability of animal models to recapitulate human disease at the molecular level. It has also been demonstrated that cellular kinases, individually or as a collective unit (the kinome), play critical roles in regulating complex biology. Despite the intimate relationship between kinases and health, little is known about the variability, consistency and stability of kinome profiles across species and individuals.

Results: As a preliminary investigation of the existence of species- and individual-specific kinotypes (kinome signatures), peptide arrays were employed for the analysis of peripheral blood mononuclear cells collected weekly from human and porcine subjects ($n = 6$) over a one month period. The data revealed strong evidence for species-specific signaling profiles. Both humans and pigs also exhibited evidence for individual-specific kinome profiles that were independent of natural changes in blood cell populations.

Conclusions: Species-specific kinotypes could have applications in disease research by facilitating the selection of appropriate animal models or by revealing a baseline kinomic signature to which treatment-induced profiles could be compared. Similarly, individual-specific kinotypes could have implications in personalized medicine, where the identification of molecular patterns or signatures within the kinome may depend on both the levels of kinome diversity and temporal stability across individuals.

9.2 Background

Efforts to increase our understanding of the mechanisms of human disease from the perspectives of both gross pathology and molecular pathogenesis have relied heavily on the use of animal models that are assumed to mimic those pathological states. Animal models, in particular those involving mice, have been employed extensively in such investigations as well as for identifying novel therapeutics and assessing their efficacy. However, many studies have relied on the similarities in the phenotypic presentation of disease rather than similarities in the underlying molecular mechanisms. Further confounding these investigations has been the assumed cross-species conservation in identities and physicochemical properties of the host molecular machinery. Although murine models have been employed extensively, there has been a relative paucity of therapeutic candidates that have translated into approved use for humans. These observations have resulted in extensive debate regarding the ability of many animal models to faithfully recapitulate human disease and to accurately predict drug efficacy in humans.

Given this, it would seem prudent to re-evaluate the criteria that drive the selection of a particular species as an animal model. Seok and colleagues recently reported that the genomic responses of mice in acute inflammatory disease models correlated poorly with those of human patients [Seok et al., 2013]. Although the authors recognized that these prior studies may have been hindered by inadequate study designs, a fatal flaw for many investigations can likely be attributed to the assumption of conservation of host responses

between mice and humans. In light of these findings, it has been suggested that a practical solution would be to select animal models based on their conservation of molecular responses to those of humans. Further, for diseases in which human clinical studies are not ethical, selection of animal models that best reflect or mimic human molecular responses would provide increased confidence in the selection or testing of therapeutics. This highlights the need for novel approaches to assess the conservation in molecular responses and identify conserved biomarkers between humans and non-human animals used in disease models.

Analyzing the conservation of molecular responses has applications not only in selecting appropriate animal models, but also in biomarker identification. While the identification and characterization of biomarkers related to disease pathology has resulted in their application to guide the diagnosis and treatment of disease [Duffy, 2001, Lilja et al., 2008], the clinical value of such biomarkers in enabling effective diagnosis or treatment guidance is dependent upon their sensitivity and specificity, which are often low [Wallis et al., 2013, Pavlou et al., 2013]. Historically, biomarkers have typically represented variations in the sequence, expression or modification of a single biomolecule. While such a simple relationship between a molecular characteristic and a phenotype is attractive from conceptual and practical perspectives, it underestimates the complexity associated with many diseases. Although some diseases are attributable to a single gene, these binary diseases represent the "low hanging fruit" of biomarker discovery. Further, in many cases diseases considered to be genetically determined have been found to display variability that must be attributed to other regulatory or phenotypic differences between individuals. Therefore, it seems appropriate to move beyond the "single gene, single disease" paradigm to a more systematic understanding of health and disease. This shift to more direct phenotypic determinants of disease often requires characterizing molecular mechanisms and biomarkers at informative, global levels. Such a systems biology approach requires examination of the dynamic interplay between large collections of biomolecules. A key challenge for the identification of biomarkers for such multi-faceted phenotypes is the development of technologies that effectively reflect these complex interactions in patient-derived samples. Investigations of dynamic patterns of gene and protein expression, through transcriptional and proteomic approaches, have offered insight into a number of disease-associated phenotypes. In cancer, for example, there are a number of biomarkers that contribute to diagnosis, subtype classification, prognosis and treatment outcomes [Gonzalez de Castro et al., 2013]. Similarly, in hepatitis C virus (HCV) infection, patterns of expression of interferon (IFN)-related genes predict IFN treatment efficacy [Barakat et al., 2012]. While these examples highlight the potential to apply global approaches to understand biology and identify biomarkers, there is concern regarding the inability of these approaches to consider post-transcriptional and post-translational regulatory events.

Kinase-mediated phosphorylation is the predominant mechanism for regulation of protein function. Disruption or dysregulation of kinase activity is associated with a number of pathophysiological states, including cancer, inflammation, neurological disorders and diabetes [Cohen, 2002]. Thus, there is considerable interest in defining kinase activities, as well as in manipulating them for therapeutic purposes—an objective facilitated by the fact that kinases are highly "druggable" [Cohen, 2002]. As a result, kinases represent a top priority

of the pharmaceutical industry [Hopkins and Groom, 2002] and are currently the most frequently targeted gene class for cancer therapies, second only to G protein-coupled receptors across all therapeutic areas [Zhang et al., 2009]. Increased appreciation for the intimate link between kinases and health has prompted the development of technologies to characterize the phosphoproteome or kinome [Jalal et al., 2007], including efforts by our group to utilize peptide arrays for low-cost, high-throughput kinomic characterizations [Turner-Brannen et al., 2011, Maddigan et al., 2011, Bodnarchuk et al., 2012, Arsenault et al., 2012, 2013b].

With respect to protein phosphorylation, the simplest evidence for the conservation (or lack thereof) of molecular responses among species lies in the content of the kinome. Genome sequencing has revealed that eukaryotes differ greatly in the number of protein kinases encoded by their genomes. For instance, the human genome encodes approximately 518 protein kinases [Manning et al., 2002]; in contrast, the proteome of *Arabidopsis thaliana* encodes around 1000 protein kinases [Champion et al., 2004], while *Saccharomyces cerevisiae* encodes fewer than 120 [Hunter and Plowman, 1997]. This suggests that kinase-mediated molecular responses may not be well-conserved among species, and that conclusions drawn from the investigation of protein phosphorylation in one species may not be applicable to another species. On the other hand, a previous study using peptide arrays suggested that despite the very different protein kinase complements in various eukaryotes, the substrates phosphorylated by these organisms exhibit substantial similarities [Diks et al., 2007]. As such, the level of conservation of kinase-mediated host responses in different species has yet to be fully delineated.

In outbred animals, it is common to observe a range of responses to a given stimulus or condition. This diversity likely reflects a combination of genetic, environmental and situational variables. Similar diversity is also apparent within human populations. In our previous investigations of livestock, unique animal-specific patterns of baseline kinome activities were often observed [Arsenault et al., 2012, 2013b]. From these animal-specific baselines, conserved yet variable responses to defined stimuli were found, suggesting that phenotypes are represented within unique cellular kinome environments. Given the close relationship between kinases and phenotype, we hypothesized that these unique signaling patterns could be used as biomarkers.

To probe the existence of species- and individual-specific kinotypes, we applied peptide arrays to conduct kinome analysis of human and porcine peripheral blood mononuclear cells (PBMCs). The peptides on the array represent phosphorylation events for which there is perfect sequence conservation between human and pig, making this array equally applicable for investigating either species. For each species, we considered six individuals sampled once per week for four consecutive weeks. The extent of conservation of kinome activity was evaluated through hierarchical clustering analysis, principal component analysis (PCA), and statistical consideration of the data. Across humans and pigs, there was overwhelming evidence for species-specific kinome profiles. The human subjects, who were variable in terms of age, gender, genetics and lifestyle, also provided evidence for individualized, stable kinome profiles. Similarly, a distinctive kinotype was observed among pigs, where potential sources of variability like age, genetics and lifestyle were minimized. The demonstration of species-specific kinotypes may have applications in the selection of animal models for

certain diseases, while the existence of stable, individualized kinotypes within members of the same species may have utility in using phosphorylation-associated biomarkers to guide disease diagnosis and treatment.

9.3 Results

9.3.1 Raw and normalized array data

For each species (human and pig), one sample was taken from each of six individuals for four consecutive weeks, for a total of 48 samples. Peptide arrays were incubated with each sample, and raw phosphorylation intensity data were collected by scanning the arrays and determining the intensity of each spot (the foreground intensity), as well as the intensity of the slide surrounding that spot (the background intensity). Because the stain binds non-specifically to the slide itself, the background intensity was often greater than the foreground intensity; in fact, only 14% of spots from the human arrays and 31% of spots from the porcine arrays had a raw foreground signal greater than the corresponding background signal. There were also differences among subjects from the same species in terms of the number of spots having a foreground signal above background. However, these systemic variations were eliminated once normalization was performed using VSN. Specifically, the average signal intensity (after background subtraction and normalization) among spots from the human arrays was 11.77 compared to 11.81 for the pig arrays, showing that measurements from the different arrays had successfully been brought onto the same scale. The raw and normalized intensity data for all arrays are available as Additional file 1 and Additional file 2, respectively.

In order to evaluate the technical reproducibility of the arrays, individual peptides (297) were printed nine times per array, and a χ^2-test was performed for each unique peptide on a given array to determine the variability amongst these technical replicates. Peptides with P-values < 0.01 were designated as inconsistently phosphorylated on that array. Over all 48 arrays, an average of 282 peptides yielded technically reproducible signals within an array (range: 268-296), giving strong evidence for the technical reproducibility of the phosphorylation signal. Due to this strong reproducibility, all 297 peptides were used for subsequent analyses.

9.3.2 Species-specific kinome profiles

Species-specific variations in phosphorylation-mediated signaling were considered as the initial test for the existence of kinotypes. Pigs were selected for comparison because they are often employed in large animal models of human diseases and therapeutic studies due to conserved biological responses and significant genetic similarities [Hein and Griebel, 2003, Groenen et al., 2012]. PBMCs were used as they are obtained through non-invasive procedures and require minimal manipulation to isolate. Further, demonstrating a kinotype within this diverse and dynamic cell population would offer confidence that individualized patterns of kinase activity would also be observed in more static and homogeneous tissues.

All human and porcine kinome profiles were analyzed simultaneously using hierarchical clustering. There

were significant differences in the profiles of humans and pigs, with nearly perfect species-specific separation of the samples (Figure 9.1a). Specifically, at the highest level of clustering, the samples separated into sample H2A (the first time point sample of human subject A) and all other samples (perhaps indicating that H2A was an outlier, as all the remaining samples for human subject A clustered exclusively with the other human samples). At the subsequent level, all remaining samples clustered into distinct, species-specific groups. To calculate the extent to which the samples clustered on the basis of species, the scoring metric T described in Materials and Methods was applied to the binary tree form of the dendrogram. The value of T was 97.9 out of 100, indicating near-perfect clustering by species. To determine whether T was greater than what would be expected by chance, the score was also calculated for 10,000 random trees. No random tree had a score > 39.6 (Figure 9.1b), giving a P-value < 0.0001. This supported the existence of species-specific patterns of kinome activity within human and porcine PBMCs.

9.3.3 Individual-specific human kinome profiles

Having demonstrated a species-specific kinotype, we investigated whether individual-specific kinomic patterns exist within members of the same species. The human subjects were investigated first as they were considered to be more likely to display significant individual differences due to variability in age, gender, race and lifestyle. Hierarchical clustering analysis revealed a clear trend for samples from the same individual to cluster together (Figure 9.2a). The score calculated for the corresponding tree was $T = 62.5$. This score was not equaled or exceeded by any of the 10,000 random trees, with the highest random tree score being 54.2, and only 0.6% of the random trees having a score > 33.3 (Figure 9.2b). This comparison again gave a P-value < 0.0001, supporting the hypothesis that individual-specific patterns of kinome activity exist within human PBMCs. The results of the clustering analysis were further verified using principal component analysis (PCA). The values of the first three principal components were calculated for each human sample and a three-dimensional scatterplot was created (Figure 9.2c). As with the hierarchical clustering, there was a strong trend for the kinome profiles to segregate on the basis of individual.

As PBMCs represent a mixed cell population, we assessed whether unique ratios of myeloid or lymphocyte subsets within an individual could be associated with particular signaling patterns. There was minimal variance between individuals with respect to the relative ratio of PBMCs over time (Table 9.1). The polymononuclear cell counts for the pigs and humans were within the normal ranges of 25-40% and 45-70%, respectively. Furthermore, there was no significant relationship between the white blood cell population variance and signaling profile variance within individuals (Figure 9.2d).

Previously, we demonstrated that monocytes purified from different animals have distinct signaling profiles [Arsenault et al., 2012, 2013a]. Thus, the differences in cell signaling profiles could reflect contributions from genetic, epigenetic or environmental variables. Although samples were collected weekly, profiles from the same individual tended to cluster together, suggesting that kinomic profiles are stable (at least over a one-month period) (Figure 9.1a). Over this time frame, there would be considerable turnover of cells and

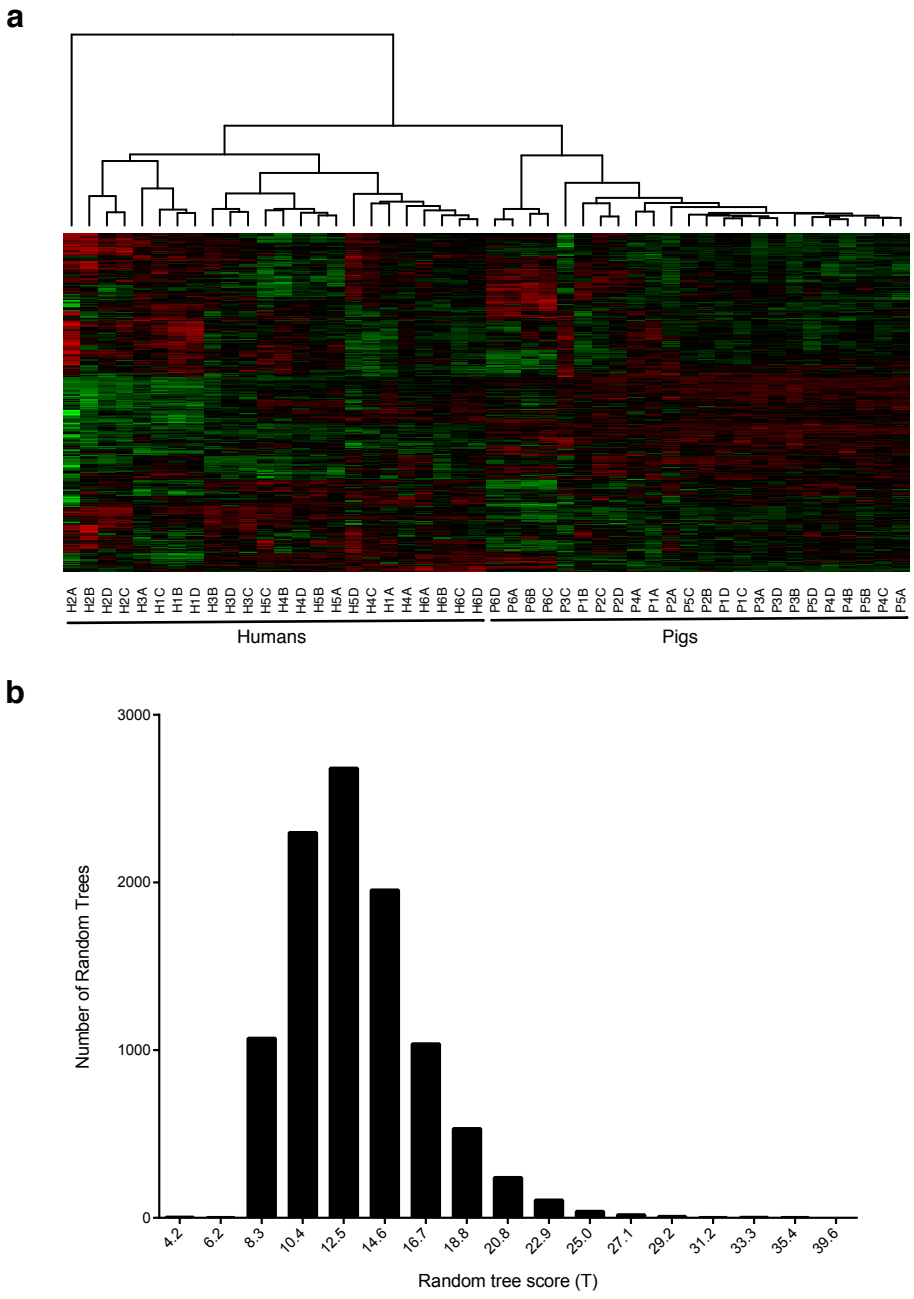


Figure 9.1: Clustering of human and porcine kinome profiles. (a) Hierarchical clustering of human and porcine kinome profiles. The distance metric used was (1 - Pearson correlation), while McQuitty linkage was used as the linkage method. Rows correspond to probes (phosphorylation targets), and columns correspond to samples. The first character of each sample label identifies the species ("H" for human and "P" for pig), the second character identifies the individual from which the sample was taken, and the third indicates the time point. Colors indicate the average (over 9 intra-array replicates) normalized phosphorylation intensity of each target, with red indicating increased phosphorylation and green indicating decreased phosphorylation. The intensity of the color corresponds to the degree of increase or decrease [Li et al., 2012]. (b) Distribution of random tree scores. The number of random trees having each random tree score is shown. For comparison, the score of the actual tree shown in part a is 97.9.

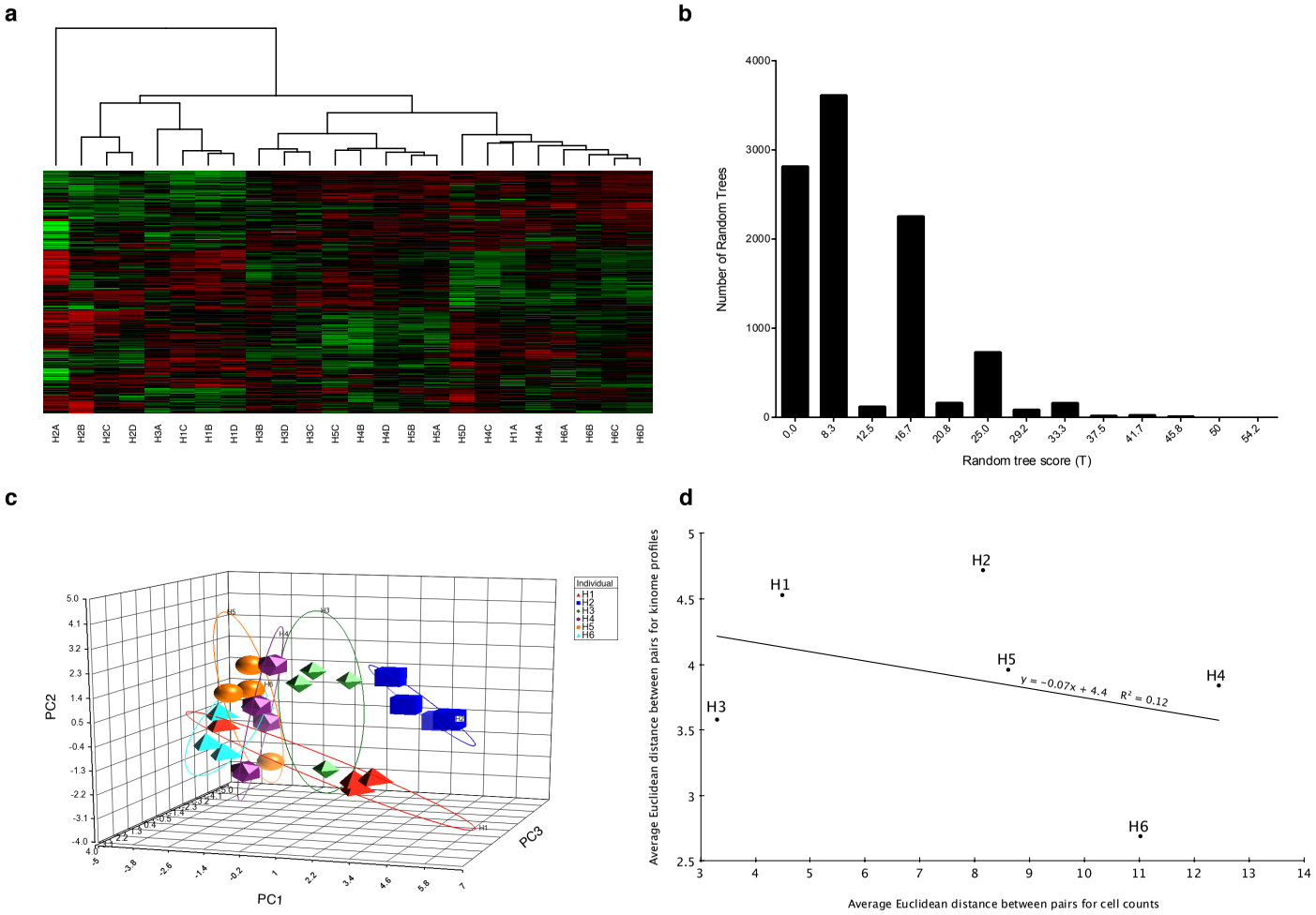


Figure 9.2: Clustering of human kinome profiles. (a) Hierarchical clustering of human kinome profiles. For details, see the caption for Figure 9.1a. (b) Distribution of random tree scores. For comparison, the score of the actual tree shown in part a is 62.5. (c) Three dimensional PCA. Individual subjects (H1, H2, H3, H4, H5, and H6) are color-coded. (d) Correlation between PBMC composition and kinome profiles. The Euclidean distance was calculated between each of the $C(4,2) = 6$ possible pairs of samples from the same individual both for cell counts (as given in Table 1) and for kinome profile (average normalized intensity values for each peptide on the corresponding array). The six Euclidean distances were then averaged for a given individual, giving a single number that represents the level of variation in either cell counts or kinome profile for that individual.

Table 9.1: Differential white blood cell counts. The first character in the sample ID column is either "H" for human or "P" for pig. The second character (e.g., "2" in "H2A") identifies the individual from which the sample was taken, while the third character (e.g., "A") indicates the time point.

Human			Pig		
ID	Lymphocytes	Monocytes	ID	Lymphocytes	Monocytes
H1A	34	10	P1A	55	12
H1B	29	10	P1B	62	1
H1C	35	9	P1C	56	5
H1D	31	6	P1D	78	5
H2A	36	7	P2A	52	8
H2B	37	9	P2B	41	10
H2C	35	2	P2C	64	1
H2D	25	5	P2D	81	8
H3A	59	3	P3A	61	4
H3B	55	5	P3B	68	3
H3C	57	4	P3C	54	5
H3D	54	6	P3D	71	2
H4A	31	7	P4A	50	7
H4B	16	6	P4B	56	5
H4C	40	4	P4C	68	4
H4D	31	5	P4D	65	4
H5A	35	7	P5A	59	7
H5B	37	6	P5B	55	4
H5C	26	8	P5C	67	7
H5D	23	8	P5D	88	5
H6A	27	8	P6A	60	4
H6B	30	6	P6B	69	0
H6C	46	5	P6C	63	3
H6D	38	6	P6D	64	5

kinases, offering some perspective on the imprinting of kinomic patterns within individuals. Several subjects displayed considerable changes in individual cell populations throughout the investigation; however, samples still clustered on the basis of individual, further supporting the existence of a stable kinotype.

9.3.4 Individual-specific porcine kinome profiles

Having demonstrated a stable, individual-specific kinome profile within human PBMCs, the concept of the kinotype was further challenged by considering the same cell population taken from pigs. In contrast to the diversity in the human subjects (gender, age, genetics, and lifestyle), the porcine subjects were littermates (siblings) housed within the same environment and sustained on the same diet. Remarkably, hierarchical clustering analysis still demonstrated a strong trend for the samples to cluster by individual ($T = 50$) (Figure 9.3a). The highest random tree score was also 50, which was achieved by a single tree (Figure 9.3b). The P-value was thus 0.0001, supporting the hypothesis that individual-specific kinotypes exist within porcine PBMCs. PCA analysis also demonstrated sample segregation on the basis of individual (Figure 9.3c). As with humans, there was no significant relationship between white blood cell population variance and kinome profile variance within individuals (Figure 9.3d).

9.3.5 Species-specific differences in the kinotypes

Having demonstrated the existence of species-specific kinome profiles of human and porcine PBMCs, we sought to identify the phosphorylation events responsible for the species-specific clustering as well as to characterize the biological events associated with them. Given that the human and porcine samples clustered separately, one would expect many peptides to be differentially phosphorylated between the two species. This was indeed the case: 119 of the 297 peptides exhibited significantly increased phosphorylation in the human samples relative to the porcine samples, while 120 peptides exhibited significantly decreased phosphorylation. Because the sample size was large for each peptide (216 observations per species), statistical significance did not necessarily imply that the difference was large in magnitude: some of the peptides with small P-values also had small fold-change (FC) values. The P-values and FC values for all peptides can be found in Additional file 3.

We have previously applied pathway over-representation analysis (ORA) to kinome data to infer cellular responses from the standpoint of signaling networks [Kindrachuk et al., 2012, Määttänen et al., 2013]. To provide initial biological insight into the observed species-specific kinotypes, here we used the Ingenuity Pathway Analysis software suite to perform functional network analysis, which provides information regarding the regulation of broad biological networks that can encompass multiple signaling pathways and cellular receptors. Differentially modulated functional networks identified from the comparison of the human and porcine profiles are presented in Figure 9.4. Functions related to cellular development, cell survival and death, and maintenance of cellular functions were over-represented, with phosphorylated mitogen-activated protein kinase (MAPK)-, signal transducer and activator of transcription (STAT)- and nuclear factor kappa-light-

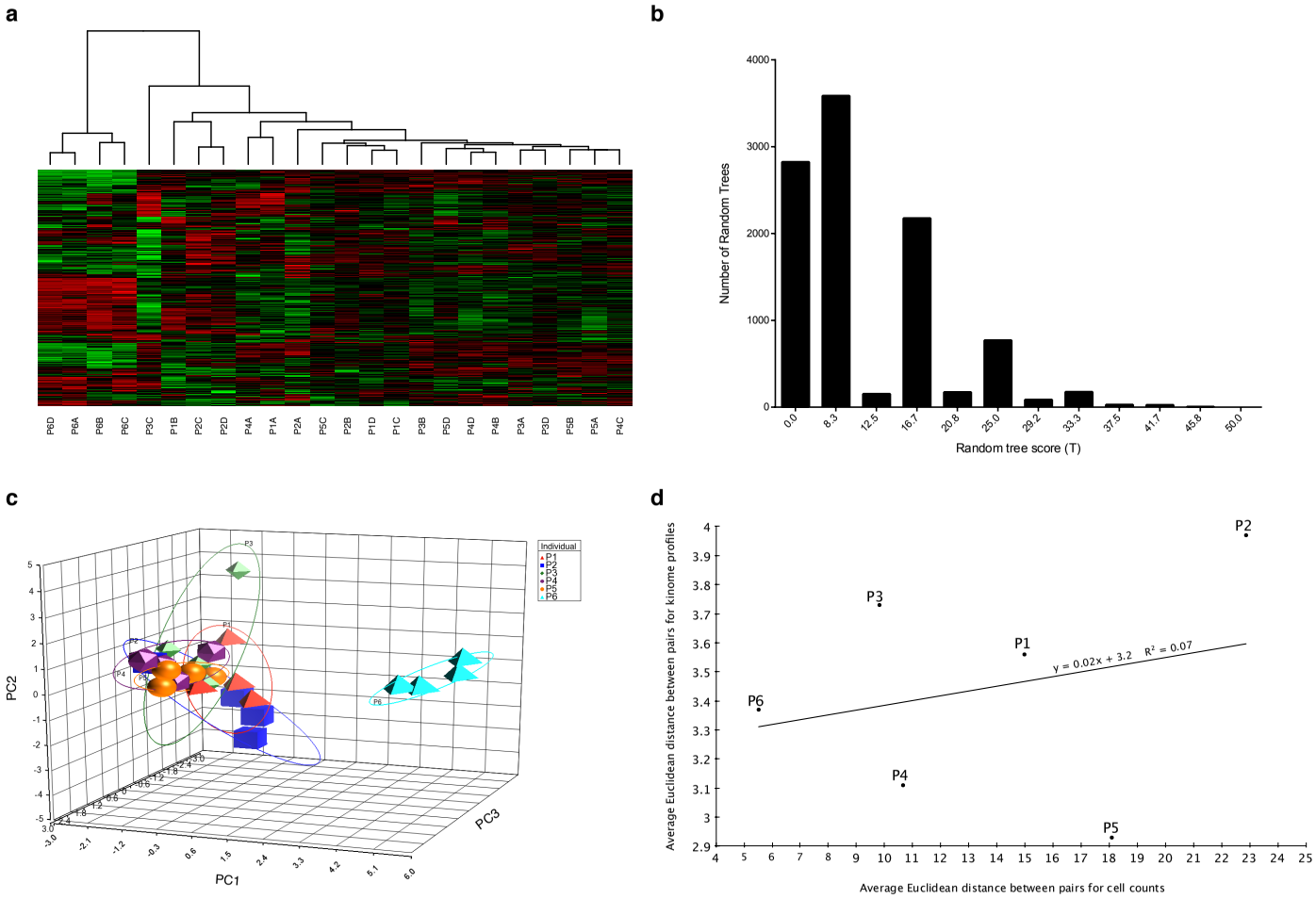


Figure 9.3: Clustering of porcine kinome profiles. (a) Hierarchical clustering of porcine kinome profiles. For details, see the caption for Figure 9.1a. (b) Distribution of random tree scores. For comparison, the score of the actual tree shown in part a is 50. (c) Three dimensional PCA. Individual subjects (P1, P2, P3, P4, P5, and P6) are color-coded. (d) Correlation between PBMC composition and kinome profiles. The procedure for generating this figure is the same as for Figure 9.2d.

chain-enhancer of activated B cells (NFκB)-regulated responses occupying central nodes of the functional network that exhibited the most significant change in modulation (Figure 9.4a). In addition, phosphorylated transforming growth factor (TGF)-β signaling pathway intermediates (including TGF-βRI and multiple SMAD proteins) formed central components of the functional network having the second-most significant change in modulation (Figure 9.4b). Additional biological verification and characterization of these kinotypic differences will be the subject of a subsequent study.

9.3.6 Individual-specific differences in the kinotypes

The individual-specific kinome profiles observed in pigs and humans support the hypothesis that kinome profiling may provide a mechanism to identify biomarkers associated with particular traits. To determine which peptides were responsible for distinguishing the kinome profiles of the individuals of a given species, the standard deviation of the normalized intensity values (averaged over 4 samples per individual and 9 technical replicates per sample) among the 6 individuals was calculated for each peptide (Additional file 4). The standard deviations of the peptides varied greatly; in human, for instance, the most variable peptide (which corresponded to the protein HSP27) had a standard deviation of 0.56, whereas the least variable peptide (IKK-α) had a standard deviation of 0.04. The range was similar in pig, with the most variable peptide (IRAK4) having a standard deviation of 0.48 and the least variable peptide (iNOS) having a standard deviation of 0.02. A moderate correlation ($r = 0.39$) was found between the standard deviation of a given peptide's response in human and the standard deviation of that peptide's response in pig, suggesting that there is some commonality between the two species in terms of the variability of the response of a given peptide among different individuals.

9.4 Discussion

Efforts to correlate phenotypes with biomolecular characteristics (e.g., nucleotide/amino acid sequences; patterns of expression/translation/modification) must often compromise between ease of technical application and biological relevance [Ludwig and Weinstein, 2005, Tan et al., 2009]. While static descriptors such as gene sequences are readily available, they often fail to capture the dynamic interplay between biological variables. In some situations, such as certain genetic disorders, the consequences associated with changes to a single biomolecule are sufficiently extreme to override this diversity. In other situations, interplay within the population of biomolecules may be of greater significance. These differences likely arise due to multiple levels of redundancy and plasticity that provide buffering for genetic differences, but also reflect individual responses. In these situations, it is most appropriate to define cellular responses at a level that reflects this interplay, ideally as close as possible to the phenotype. The challenge here is that unlike genetic polymorphisms, the levels of these biomolecules are dynamic and may be unique to particular tissues, cells or intracellular locations. From a practical perspective, there is also the need to be able to reliably quantify

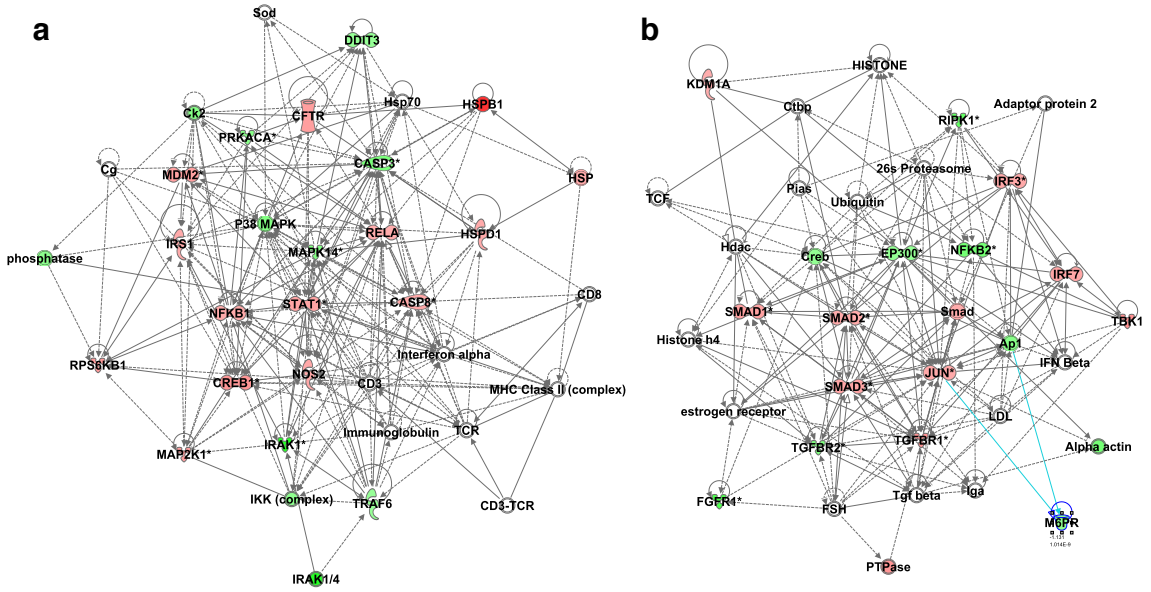


Figure 9.4: Functional network analysis of differentially modulated kinome responses in humans as compared to pigs. Peptide fold-change values for humans as compared to pigs were uploaded to IPA for functional network analysis. Differentially modulated functions were identified as follows: (A) cellular development, cell death and survival, cellular function and maintenance; and (B) gene expression, digestive system development and function, cell death and survival. As in the heatmaps, red indicates increased phosphorylation and green indicates decreased phosphorylation.

these biomolecules in a robust, cost-effective and high-throughput manner.

Protein kinases play a central role in regulating biological functions at the levels of proteins through to pathways and, ultimately, phenotypes. Changes in activities of individual kinases, through genetic defects or therapeutic modulation, can have profound impacts on the health and viability of an organism [Graves et al., 2013]. The growing interest in kinases in both basic and translational research has driven efforts to develop technologies that enable characterization of phosphorylation-mediated signal transduction [Jalal et al., 2009, Harsha and Pandey, 2010]. Kinome analysis offers three key advantages over traditional gene and protein profiling: 1) individual kinase activities are often reliable indicators of phenotypic changes, 2) kinase profiling offers insight into cellular responses at the level of signaling pathways, and 3) as kinases are highly "druggable" [Cohen, 2002, Hopkins and Groom, 2002], increased understanding of the biological role for kinases could aid therapeutic design and development.

To this end, our investigation sought to examine kinome responses in both inter- and intra-species comparatives. Our results demonstrate the existence of temporally stable species- and individual-specific kinotypes. Hierarchical clustering of the kinome data derived from human and porcine PBMCs showed that the kinome profiles clustered in a species-specific manner, suggesting that kinotype analysis could provide critical information regarding cell processes or signaling pathways that are differentially modulated across similar animal species.

In addition to verifying the existence of species-specific kinotypes, this finding may have further implications and applications. Recently, Seok and colleagues reported on the disparate correlation in genomic responses between human inflammatory diseases and murine inflammatory models [Seok et al., 2013]. This highlights a problem in basic and translational research where animal models of disease are largely validated through phenotypic similarity to human disease. Our results suggest that kinome analysis could be beneficial for the evaluation and assessment of animal models. Indeed, pigs are often employed as animal models of human disease. Thus, understanding the biological differences or predispositions between pigs and humans could inform situations where pigs may represent an appropriate animal model as well as influence the interpretation of emerging results. Further, our demonstration of species-specific kinotypes suggests that kinotype analysis could provide critical information regarding the degree of conservation of essential cell processes across animal species, in particular those that are closely related. Further, we postulate that kinome analysis could provide information regarding conserved mechanisms of molecular pathogenesis between humans and animals routinely used in models of human malignancies. With respect to the current US Food and Drug Administration Animal Efficacy Rule [US Department of Health and Human Services, 2002, Gronvall et al., 2007], kinome analysis could provide insight in investigations for which human efficacy trials are neither feasible nor ethical and, in particular, in the selection of animal models that best recapitulate human molecular disease.

From the standpoint of drug development, the analysis of individual-specific kinotypes could help define the temporal stability of particular drug targets as well as their conservation across the population. Personal-

ized medicine is based on an appreciation that natural biological variation exists within outbred populations. Customizing diagnoses and therapies to an individual rather than an assumed biological norm has the potential to maximize treatment efficacy while minimizing side effects [Chiang and Million, 2011, Mehta et al., 2011]. The implementation of personalized medicine at a molecular level depends on the identification of biomarkers that accurately predict some aspect of disease, such as onset, prognosis or treatment efficacy. It is our belief that kinotype analysis could facilitate this process.

Beyond this study, there is a substantial opportunity for future work in terms of expanding the number of model organisms considered. Perhaps the most important information that could be derived from an analysis involving several species would be to determine which has a kinome profile most similar to that of human. More generally, it would be interesting to compare the clustering of the different species' kinome profiles with those obtained from traditional sequence-based phylogenetic approaches (e.g., mitochondrial 16S rRNA gene comparisons). Answering the question, "Is there a strong relationship between genetic similarity and kinotypic similarity?" would be hugely beneficial in terms of selecting appropriate animal models and understanding how well the responses of a given model might reflect those in human. Another avenue for future work derives from the fact that the number of samples could become quite large if several species are considered, especially if many individuals are tested per species and/or many samples are taken per individual. In this study, four samples were taken from each of six individuals from each of two species, for a total of 48 samples. In addition, each sample was exposed to a peptide array with nine intra-array technical replicates per peptide sequence. Kinome microarrays are relatively inexpensive; nonetheless, in order to provide accurate comparisons while simultaneously minimizing costs, it could be beneficial to characterize the number of intra-array replicates per sample, samples per individual, and individuals per species required to accurately reflect the level of kinotypic similarity among species and individuals.

9.5 Conclusions

The identification of phosphorylation signatures associated with disease states using kinome analysis may become an important tool in basic and translational research. This study suggests that these signatures must be considered in the context of the range of variability at the level of both species and individual. For instance, if an animal is being considered as a model for a particular disease—and in particular, if host responses are being evaluated at the kinome level—then species-specific baseline levels of kinase activity may need to be taken into account. The same concept applies in the context of personalized medicine: a treatment that is effective in some individuals may not be effective in other individuals, and it is possible that an individual's kinome profile may be predictive of the efficacy of a given treatment. Of course, further studies are needed to precisely define the methodology needed for incorporating kinome analysis into both treatment studies and studies involving animal models. While considering baseline kinomic responses may prove complicated, the discovery of complex biomarkers, in particular those associated with kinase activities,

has tremendous potential to inform research involving animal models as well as personalized medicine.

9.6 Materials and methods

9.6.1 PBMC isolations

Human and porcine blood samples were collected weekly for 4 consecutive weeks. For humans, 6 unrelated individuals (3 male and 3 female) diverse in age (21-55), race, diet, and health status were selected. Porcine samples were obtained from 6 littermates (3 male and 3 female) beginning at 4 weeks of age. Pigs were housed within the same pen and fed the same diet. Bleeds were performed at the same time each day to minimize variability associated with circadian rhythms and postprandial effects. PBMCs were isolated as previously described [Kindrachuk et al., 2011]. Aliquots of 1×10^7 PBMCs were snap-frozen in liquid N_2 and stored at $-80\,^{\circ}C$ for kinome analysis. All animal procedures were performed in accordance with the standards of the Canadian Council on Animal Care. Human subjects provided written informed consent before participation. This procedure, and all research done using these samples, was done in accordance with the University of Saskatchewan Clinical Research Ethics Board.

9.6.2 Peptide arrays

Design, construction and application of the peptide arrays were based on previous protocols with modifications [Jalal et al., 2009]. A commercial provider, JPT Peptide Technologies (`http://www.jpt.com`), was contracted to fabricate the arrays. Peptides from proteins representing a wide variety of signaling pathways were included on the arrays. Specifically, 297 peptide sequences were chosen, each of which was spotted 9 times on the same array (i.e., there were 9 intra-array technical replicates per peptide). It should be noted that this type of technical replicate is distinct from inter-array technical replicates, for which the entire process (from incubation of the sample with the array to scanning the array using image analysis software) is repeated multiple times. The technical replicates for a given peptide were averaged prior to performing clustering analysis. The exact composition of the array, including spot coordinates, block layouts, and peptide sequences, is given in Additional file 5.

All peptides on the array are found as exact matches in both the human proteome and the porcine proteome; as such, the same arrays were used for both species, enabling a direct comparison of all kinome profiles. Kinome array experiments for both species were performed on the same day to minimize technical variance and performed as described previously [Booth et al., 2010]. Each resulting dataset contained the signal intensities associated with all 9 replicates of the 297 peptides from a given individual at a given time point. All data processing and analysis was done using the Platform for Intelligent, Integrated Kinome Analysis (PIIKA) software [Li et al., 2012], which is freely available for non-commercial use at `http://saphire.usask.ca/saphire/piika`.

9.6.3 Evidence for individual kinotypes in humans and pigs

To determine whether unique kinotypes existed within individuals, the following statistical question was addressed: "Do samples from the same individual cluster more closely than expected by chance?" Samples from the 6 individual humans and pigs were separately subjected to hierarchical clustering using (1 - Pearson correlation) as the distance metric and McQuitty linkage as the linkage method. We defined a metric describing how close to perfect (i.e., all samples from the same individual clustering together) the actual clustering was. As each iteration of the hierarchical clustering algorithm results in a bifurcation, the resulting dendrogram can be represented as a binary tree wherein each leaf represents one of the 24 samples, and each internal node represents a cluster of 2 or more samples. For each individual i, a score s_i was computed. If some internal node in the binary tree had, as descendants, four leaves corresponding to individual i and none corresponding to any other individual, then $s_i = 4$. If the same criteria could be satisfied but with only 3 leaves corresponding to individual i, then $s_i = 3$, and similarly for $s_i = 2$ and $s_i = 1$. If there were no internal nodes having, as descendants, only leaves corresponding to individual i, then $s_i = 0$.

The score for the entire tree was $S = \sum_{i=1}^{6} s_i$, with the maximum possible score being 24. This was then expressed as a score out of 100: $T = S/24 \times 100$. To determine whether T was greater than would be expected by chance, an empirical statistical distribution was derived by generating 10,000 random trees. Each tree was created by randomly rearranging the normalized intensity values for the peptides within a given sample. The average normalized intensity value (over the 9 technical replicates) for a given peptide X was randomly assigned to a different peptide Y from the same sample, and this was done for all peptides across all samples. For each random tree j, hierarchical clustering was performed and a score T_j was calculated as described above. The P-value for a given score T was then calculated as the proportion of scores T_j that were $\geq T$.

9.6.4 Evidence for species-specific kinotypes

To answer the question, "Do samples from the same species cluster more closely than expected by chance?", hierarchical clustering was performed with all samples from both species at once. The same scoring metric as above was used, but with only 2 "individuals"—human and pig, each with 24 samples. Thus, $S = \sum_{i=1}^{2} s_i$, with s_1 and s_2 denoting the scores for the human and porcine samples, respectively, and $T = S/48 \times 100$. Statistical significance was determined as above.

9.6.5 Correlating cell composition and kinome profiles

Blood contains a dynamic population of cells that, based on their unique functions, likely exhibit distinct signaling activity. Thus, species- or individual-specific kinome patterns could reflect unique blood cell populations. To account for this potential variability, differential counts were performed on each sample. Importantly, kinome analysis was performed solely on PBMCs (lymphocytes and monocytes) and excluded polymononuclear cells (PMNs). We investigated whether there was any correlation between kinome profiles

and relative abundance of PBMCs as follows.

The level of variability within an individual's kinome profile over time was determined by finding the Euclidean distance between each of the 6 possible pairs of samples for the same individual (week 1 and week 2, week 1 and week 3, etc.) with respect to the average normalized intensities for each peptide. Specifically, each sample was represented as a vector of length 297, where each element represented the average normalized intensity value for a peptide on the array. For a given pair of samples x and y, the Euclidean distance was calculated as $\sqrt{\sum_{i=1}^{297}(x_i - y_i)^2}$. The level of variability in a given individual's kinome profile was the average of all 6 Euclidean distances. The level of variability in cell counts over time was assessed similarly, except the values of a given vector represented counts for a given cell type. As such, these vectors were of length 2 (lymphocytes and monocytes). To determine whether there was a relationship between the variables mentioned above, a scatterplot was created for each species wherein the independent axis represented variability in cell counts for a given individual, and the dependent axis represented variability in kinome profile. Linear regression was performed, and the coefficients of the regression line and the R^2 value were calculated for each species.

9.6.6 Species-specific differences in the kinotypes

Statistical tests for identifying peptides differentially phosphorylated in the human samples compared to the porcine samples were carried out as described previously [Li et al., 2012]. Specifically, for each peptide, a t-test was done by comparing all 216 human observations (6 subjects $\times$ 4 samples per subject $\times$ 9 technical replicates per sample) against all 216 porcine observations. A peptide was considered to be differentially phosphorylated if the resulting P-value was less than 0.05. Pathway overrepresentation analysis was performed as previously described [Li et al., 2012], except that the Ingenuity Pathway Analysis software suite was used instead of InnateDB.

9.6.7 Individual-specific differences in the kinotypes

In order to identify peptides driving the differences between the kinome profiles of different individuals, the 36 normalized intensity values for a given individual (4 samples per individual $\times$ 9 technical replicates per sample) were averaged for each peptide. Within each species, the standard deviation of these values for the 6 individuals was calculated. Peptides with high standard deviation had the greatest variation in responses among the individuals, while peptides with low standard deviation had the most consistent responses.

9.7 Acknowledgements

This work was supported in part by Genome Canada, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Intramural Research Program of the NIH/NIAID.

9.8 Additional files

Additional file 1: Raw array intensity data for all human and porcine samples (raw_intensity_data.xlsx)—Columns correspond to samples and rows correspond to peptides. The rows are in groups of 9, representing the values for the 9 technical replicates associated with a given peptide. The first column for a given sample represents the foreground intensity, while the second column represents the background intensity.

Additional file 2: Intensity data after background subtraction and normalization (normalized_intensity_data.xlsx)—Columns correspond to samples and rows correspond to peptides. Unlike Additional file 1, there is only one row corresponding to a given peptide; however, there are 9 columns for each array, which give the normalized intensity values for the 9 technical replicates for that peptide.

Additional file 3: Comparison of human and porcine kinome responses (human_pig_comparison.xlsx)—The fold-change value between human and pig is given for each peptide on the array, along with P-values for increased and decreased phosphorylation.

Additional file 4: Inter-individual variability of peptide responses (individual_variability.xlsx)—The left-hand block of cells contain, for a given peptide and individual, the mean normalized intensity value among the 36 observations (4 samples per individual and 9 technical replicates per sample). Column O contains the standard deviation of the 6 human means, while column R contains the standard deviation of the 6 porcine means.

Additional file 5: Composition of the peptide arrays (peptide_array_composition.gal)—This GenePix Array List (GAL) file contains the exact composition of the peptide array used in this study, including the location of each spot and the peptide contained in that spot. It is in plain-text format and can thus be read by any text editor.

Chapter 10

Divergent immune responses to *Mycobacterium avium* subsp. *paratuberculosis* infection correlate with kinome responses at the site of intestinal infection

Pekka Määttänen, Brett Trost, Erin Scruten, Andrew Potter, Anthony Kusalik, Philip Griebel, and Scott Napper

This is the second of three papers that describe biological applications of the work described in this thesis. It describes the application of kinome arrays (as well as other biological techniques) to the study of *Mycobacterium avium* subsp. *paratuberculosis* (MAP) infection in calves, which can cause Johne's disease (JD). Kinome microarray analysis was performed on tissue samples taken from calf intestinal segments, and it was found that the kinome profile of a given segment was related to the effectiveness of the immune response generated by the calf. This information could aid in the discovery of a treatment for JD, which could employ the general strategy of guiding the immune response to the kind that is effective against JD.

Citation

P. Määttänen, B. Trost, E. Scruten, A. Potter, A. Kusalik, P. Griebel, and S. Napper. Divergent immune responses to *Mycobacterium avium* subsp. *paratuberculosis* infection correlate with kinome responses at the site of intestinal infection. *Infect Immun* 81(8):2861-2872, 2013.

Author contributions

Pekka Määttänen designed the experiments, performed the majority of the experiments, analyzed most of the data, and wrote the majority of the paper. Brett Trost performed the analysis of the kinome microarray data, wrote the corresponding portions of the paper, helped with other aspects of data analysis, and revised the paper. Erin Scruten performed the kinome microarray experiments. Andrew Potter supervised the research. Anthony Kusalik supervised the research and revised the paper. Philip Griebel and Scott Napper helped design the experiments, supervised the research, and revised the paper.

Supplementary material

Supplementary material for this paper are given in Appendix F.

10.1 Abstract

Mycobacterium avium subsp. *paratuberculosis* is the causative agent of Johne's disease (JD) in cattle. *M. avium* subsp. *paratuberculosis* infects the gastrointestinal tract of calves, localizing and persisting primarily in the distal ileum. A high percentage of cattle exposed to *M. avium* subsp. *paratuberculosis* do not develop JD, but the mechanisms by which they resist infection are not understood. Here, we merge an established *in vivo* bovine intestinal segment model for *M. avium* subsp. *paratuberculosis* infection with bovine-specific peptide kinome arrays as a first step to understanding how infection influences host kinomic responses at the site of infection. Application of peptide arrays to *in vivo* tissue samples represents a critical and ambitious step in using this technology to understand host-pathogen interactions. Kinome analysis was performed on intestinal samples from 4 ileal segments subdivided into 10 separate compartments (6 *M. avium* subsp. *paratuberculosis*-infected compartments and 4 intra-animal controls) using bovine-specific peptide arrays. Kinome data sets clustered into two groups, suggesting unique binary responses to *M. avium* subsp. *paratuberculosis*. Similarly, two *M. avium* subsp. *paratuberculosis*-specific immune responses, characterized by different antibody, T cell proliferation, and gamma interferon (IFN-γ) responses, were also observed. Interestingly, the kinomic groupings segregated with the immune response groupings. Pathway and gene ontology analyses revealed that differences in innate immune and interleukin signaling and particular differences in the Wnt/β-catenin pathway distinguished the kinomic groupings. Collectively, kinome analysis of tissue samples offers insight into the complex cellular responses induced by *M. avium* subsp. *paratuberculosis* in the ileum and provides a novel method to understand mechanisms that alter the balance between cell-mediated and antibody responses to *M. avium* subsp. *paratuberculosis* infection.

10.2 Introduction

Johne's disease (JD) of cattle and other ruminants is caused by a chronic enteric infection by *Mycobacterium avium* subsp. *paratuberculosis*. JD is characterized by a long asymptomatic latency period during which animals display variable humoral and inflammatory immune responses [Waters et al., 2003, Wu et al., 2007]. During the symptomatic phase of infection, there is a progressive inflammatory enteritis, diarrhea, and significant weight loss [Sweeney et al., 2012]. The infected host sheds *M. avium* subsp. *paratuberculosis* throughout the course of infection but especially during the late stages. Shedding occurs primarily in the feces [Marcé et al., 2011] but has also been detected in milk [Giese and Ahrens, 2000]. Large quantities of *M. avium* subsp. *paratuberculosis* shed by infected cattle survive for extended periods in the environment and persist after high-temperature, short-time pasteurization, raising concerns about contamination of dairy products [Ellingson et al., 2005]. *M. avium* subsp. *paratuberculosis* has been shown to infect primates and has been postulated to be a cause of Crohn's disease in humans [Sweeney et al., 2012]. The zoonotic potential of this infection and its already devastating impact on cattle and sheep have fueled extensive study into its

pathogenesis. A priority is to determine the mechanisms by which *M. avium* subsp. *paratuberculosis* subverts the host immune system to establish chronic infection. Understanding these mechanisms may represent a critical step toward the development of either effective vaccines or therapeutics.

Cattle exhibit highly variable responses to *M. avium* subsp. *paratuberculosis* infection, and, similar to the well-documented high rates of human resistance to *Mycobacterium tuberculosis* [Kleinnijenhuis et al., 2011], not all calves exposed to the pathogen develop JD [Koets et al., 2000]. This suggests that genetic and/or environmental factors predispose animals to disease. A recent meta-analysis of two genome-wide association studies revealed multiple loci associated with *M. avium* subsp. *paratuberculosis* infection of cattle, indicating that genetic susceptibility to infection is complex, involving 11 different chromosomes [Minozzi et al., 2012]. Therefore, understanding the regulation of immune responses to this pathogen in an outbred population is a daunting challenge. Furthermore, the eradication of *M. avium* subsp. *paratuberculosis* is complicated by its persistence in soil, feed, and water, and as a result, strategies for controlling the spread of infection have focused on herd management. The ability to distinguish animals that effectively control *M. avium* subsp. *paratuberculosis* infection from more susceptible animals might provide a way to selectively enhance the health of cattle and decrease the zoonotic threat.

Monitoring global responses at the level of cellular kinase activity (the kinome) is an effective approach to understand complex biology as well as to identify therapeutic targets and biomarkers [Arsenault et al., 2011]. Many methods have been used to assay kinase activity under different conditions, each with advantages and disadvantages [Knight et al., 2013]. While peptide array approaches that attempt to identify novel phosphorylation sites may lead to false positives, focused arrays that employ a subset of better-characterized phosphorylation sites are powerful tools to profile pathways of interest. Specifically, kinome profiling offers a way to differentiate individuals at a phenotypic level, with the potential to reveal adaptive or maladaptive shifts in host signaling patterns in response to a pathological state such as infection. Previously, using a bovine-specific peptide array developed in our lab, we used kinome analysis to reveal specific mechanisms through which *M. avium* subsp. *paratuberculosis* influences the ability of bovine monocytes to respond to gamma interferon (IFN-γ) and Toll-like receptor ligands [Arsenault et al., 2012, 2013a]. These investigations highlighted mechanisms employed by the pathogen to alter innate immune responses as well as the power of kinomics to reveal host signaling events following infection.

Here, employing a bovine intestinal segment model developed by our group to restrict *M. avium* subsp. *paratuberculosis* infection to specific sites in the intestine [Charavaryamath et al., 2013], we monitored adaptive immune responses in parallel with kinome profiling of ileal tissues from *M. avium* subsp. *paratuberculosis*-infected and uninfected intestinal compartments. Kinome profiling of tissues is an ambitious but not unprecedented approach to understand shifts in kinase activities in pathology, and it has provided insights into aberrant signaling in different cancers [Kilpinen et al., 2010, Grzmil et al., 2011, Hildebrandt et al., 2012]. We targeted the early (1-month) phase of *M. avium* subsp. *paratuberculosis* infection since antibody responses can be detected in subclinical *M. avium* subsp. *paratuberculosis* infections [Waters et al., 2003].

Furthermore, it has been suggested that host responses during the first few weeks after infection may determine whether JD develops [Whittington et al., 2012]. We hypothesized that *M. avium* subsp. *paratuberculosis* infection for 1 month should provide sufficient time for imprinting the host immune response to *M. avium* subsp. *paratuberculosis* and allow us to explore potential relationships between global kinase activity at the site of infection and immune responses occurring in gut-associated lymphoid tissue.

10.3 Materials and methods

10.3.1 Calves, surgery, and infection with *Mycobacterium avium* subsp. *paratuberculosis*

All experimental protocols were performed following the guidelines approved by the Canadian Council on Animal Care. Protocols for animal housing, anesthesia, surgery, *M. avium* subsp. *paratuberculosis* infection, and postsurgical care were performed as previously described [Charavaryamath et al., 2011, 2013]. Young calves are most susceptible to JD [Marcé et al., 2011], and four calves that were 2 weeks old were inoculated with *M. avium* subsp. *paratuberculosis* in surgically isolated intestinal compartments. Briefly, a 30- to 35-cm segment of intestine was surgically isolated, proximal to the ileocecal fold, and subdivided into three equal compartments using silk ligatures. The distal and middle compartments were injected with 1×10^8 to 3×10^8 CFU of *M. avium* subsp. *paratuberculosis* strain K10 (preparation is described below) in a final volume of 5 ml phosphate-buffered saline (PBS). The proximal compartment of the intestinal segment was injected with 5 ml PBS. Postsurgically, calves were treated with 1.1 mg/kg flunixin (Banamine; Schering Plough Canada Inc., Pointe Claire, Quebec, Canada) for 3 days and with 3 to 4 mg/kg enrofloxacin (Baytril; Bayer Inc.) for 5 days. Three uninfected calves of similar age from the same herd were included as negative controls. Calves were maintained on a whole-milk diet for 4 weeks. Blood for the isolation of serum and peripheral blood mononuclear cells (PBMCs) was collected 1 day before animals were euthanized for collection of tissue samples (ileal intestinal compartments and contents, spleen, and mesenteric lymph node [MLN]).

10.3.2 Preparation of *M. avium* subsp. *paratuberculosis* inoculum and lysate

M. avium subsp. *paratuberculosis* strain K10 was a generous gift from Marcel Behr (McGill University Health Center, Montreal, Quebec, Canada). One loop of K10 culture grown on Middlebrook 7H10 agar (BD Bioscience, Canada) was inoculated into 50 ml of Difco Middlebrook 7H9 broth (BD Bioscience, Canada) and incubated on a rotary shaker for 5 days at 37 °C. *M. avium* subsp. *paratuberculosis* cultures were centrifuged at $3,000 \times g$ at 4 °C for 15 min in a preweighed sterile 50-ml centrifuge tube. The supernatant was discarded, and the bacterial pellet was allowed to dry for 30 min in the inverted tube before obtaining a final weight. The weight of the dry bacterial pellet (tube with pellet minus tube weight) was calculated and total CFU calculated on the basis of previous titration experiments [Charavaryamath et al., 2013]. *M. avium* subsp.

paratuberculosis lysate for cell stimulation experiments was prepared from 5-day cultures of *M. avium* subsp. *paratuberculosis* grown in Middlebrook 7H9 broth as described previously [Charavaryamath et al., 2013] except that the protein concentration of the lysate was measured using the Bio-Rad Bradford reagent (Bio-Rad Laboratories, Mississauga, Ontario, Canada). Phenylmethylsulfonyl fluoride (PMSF) (Sigma-Aldrich, Canada) was added to the bacterial lysate (final concentration, 1 mM) before storage at $-20\,^{\circ}$C.

10.3.3 Tissue collection and histology

Tissues for kinome analysis and pathology were collected from infected and uninfected intestinal compartments immediately after euthanizing the calves. A 4- to 5-cm segment of each infected ileal compartment and the adjacent uninfected ileal compartment, including contents, was collected and fixed in 10% neutral buffered formalin (VWR, Westchester, PA) for histopathological examination. The remaining tissue from each compartment was then opened longitudinally, the contents were removed, and longitudinal strips of intestinal tissue measuring 0.5 cm by 3 cm were collected, placed into cryovials, and snap-frozen in liquid nitrogen prior to storage at $-80\,^{\circ}$C. Histology samples fixed in 10% neutral buffered formalin (NBF) were embedded, sectioned, and acid-fast stained by Prairie Diagnostic Services (Saskatoon, Saskatchewan, Canada). Photos were taken through an Olympus BX41 microscope with a flip-out condenser and a 100× UPlan Fluorite oil immersion lens using a 12 Megapixel Olympus DP71 camera with DP controller acquisition and managing software (Olympus).

Blood was collected from the jugular veins of calves at 4 weeks after *M. avium* subsp. *paratuberculosis* infection and 1 day prior to euthanizing. PBMCs were isolated following a previously described protocol [Whale et al., 2006]. Briefly, blood was centrifuged at $1,400 \times g$ for 20 min at room temperature before collecting the buffy coat and resuspending cells in 35 ml Ca^{2+}- and Mg^{2+}-free PBS (PBSA) containing 0.1% EDTA. Cells were layered onto isotonic 54% Percoll (GE Health Bio-Sciences AB, Uppsala, Sweden) and centrifuged at $2,000 \times g$ for 20 min at room temperature. Cells at the Percoll-PBS interphase were collected and washed three times with PBS before resuspending PBMCs at a final concentration of 2×10^6 viable cells/ml in RPMI medium (Invitrogen, Burlington, Ontario, Canada) containing 10% fetal bovine serum (FBS) (Invitrogen) plus antibiotics and antimycotics (Sigma-Aldrich Canada, Oakville, Ontario, Canada). Mesenteric lymph nodes (MLNs) and spleens were collected immediately after euthanizing calves, and tissues were placed in ice-cold Dulbecco modified Eagle medium (DMEM) (Sigma-Aldrich Canada). For lymph nodes, pericapsular fat was removed before the lymph node was cut and immersed in PBSA. The tissue was minced finely with a scalpel blade to release single cells. The cell suspension was passed through a 40-μm nylon cell strainer (Becton Dickson Labware, Franklin Lakes, NJ) and cells washed three times with PBSA containing 0.1% EDTA and one time with only PBSA before being resuspended at a final concentration of 2×10^6 cells/ml in DMEM supplemented with 10% FBS, antibiotics, and antimycotics. Spleens were minced and strained as described above to release and isolate single cells, and the cells were centrifuged at 1,200 rpm for 10 min at 4 °C. Medium was poured off and red blood cells lysed by a short treatment with double-distilled water

(ddH_2O), followed by addition of 1/10 volume of 10× PBSA to restore isotonic pH and three washes in PBSA. Splenocytes were counted on a hemocytometer using trypan blue stain and resuspended to a concentration of 2×10^6 cells/ml in DMEM supplemented with 10% FBS, antibiotics, and antimycotics.

10.3.4 Immune assays

PBMCs, MLN cells, and splenocytes were cultured in 96-well tissue culture plates (Thermo Fisher Scientific, Rochester, NY) in a final volume of 200 μl DMEM supplemented with 5% FBS, antibiotics, and antimycotics. Separate 96-well plates were set up to allow two separate stimulations, one for measurement of lymphocyte proliferation (2×10^5 cells/well in triplicate) and the other to quantify IFN-γ secretion (5×10^5 cells/well in duplicate). For each lymphoid tissue, cultures were stimulated with either medium alone, 1 μg/ml *M. avium* subsp. *paratuberculosis* lysate, or 1 μg/ml concanavalin A (Sigma-Aldrich Canada), and the cultures were incubated at 37 °C in a humidified atmosphere containing 5% CO_2. IFN-γ secretion into the culture supernatants was measured 48 h after stimulation with a capture enzyme-linked immunosorbent assay (ELISA), as described previously [Raggo et al., 2000]. Proliferation assay culture plates were incubated for 5 days. During the last 6 h of culture, 20 μl [^{3}H]thymidine (GE Healthcare Biosciences, Pittsburg, PA) (0.4 μCi per well) was added to each well. Plates were freeze-thawed to lyse cells and harvested using a Microplate cell harvester (model C961961; Perkin-Elmer, Waltham, MA). Radioactivity was detected using a Top Count NXT beta scintillation counter (model C9912V1; Perkin-Elmer, Waltham, MA). Average counts per minute (cpm) were calculated for each set of triplicate cultures, and stimulation indices (SIs) were calculated by dividing average cpm for cells stimulated with *M. avium* subsp. *paratuberculosis* lysate by average cpm for cells cultured in medium alone.

10.3.5 Immunoblotting

Total *M. avium* subsp. *paratuberculosis* lysate prepared in PBS containing protease inhibitors as described above was supplemented with concentrated Laemmli SDS-PAGE buffer to 1× with a final concentration of 2.5% β-mercaptoethanol. The lysate was heated for 5 min at 95 °C and then cooled to room temperature. Lysate samples (1.25 μg/lane) were separated by SDS-PAGE in a 1.0-mm Tris-glycine minigel and blotted onto a polyvinylidene difluoride (PVDF) membrane. Membranes were blocked with 5% skim milk-Tris-buffered saline (TBS)-0.05% Tween 20. Bovine sera collected prior to and after experimental infection were diluted 1:100 in 5% skim milk-TBS-Tween 20 before being applied to each blot. Blots were washed, and bound antibody was detected with alkaline phosphatase (AP)-conjugated goat anti-bovine IgG(H+L) antibody (KPL Laboratories, Gaithersburg, MD), followed by SigmaFast 5-bromo-4-chloro-3-indolylphosphate (BCIP)-nitroblue tetrazolium (NBT) (Sigma-Aldrich, St. Louis, MO). Blots were dried and scanned with an HP Scanjet G4050 at 600 dpi (grayscale).

10.3.6 Kinome array experiments

The design, construction, and application of the bovine peptide arrays were carried out essentially as previously described [Jalal et al., 2009, Arsenault et al., 2009, 2013a]. The kinome arrays used were specifically designed for analysis of signaling involved in immune function and were previously applied to bovine monocytes infected with *Mycobacterium avium* subsp. *paratuberculosis* [Arsenault et al., 2012, 2013a]. Kinome array experiments were all performed within a single assay on the same day to minimize technical and interassay variance. Briefly, ileal intestinal tissue samples flash-frozen and stored at $-80\,^{\circ}\mathrm{C}$ were crushed in liquid nitrogen using a ceramic mortar and pestle. Pulverized tissue in liquid nitrogen was transferred to a tared sterile 1.5-ml microcentrifuge tube, the liquid nitrogen was allowed to evaporate, and the tissue was weighed. An 80-μl aliquot of ice-cold lysis buffer (20 mM Tris-HCl [pH 7.5], 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton, 2.5 mM sodium pyrophosphate, 1 mM Na_3VO_4, 1 mM NaF, 1 μg/ml leupeptin, 1 μg/ml aprotinin, 1 mM PMSF, 1 μg/ml pepstatin A, and 2 mM dithiothreitol [DTT] [all products from Sigma-Aldrich unless otherwise indicated]) was added per 15 mg of tissue and vortexed to mix. This homogenate was further diluted 8-fold in ice cold lysis buffer, vortexed thoroughly, and incubated on ice for 20 min to allow adequate lysis. Lysates were spun in a microcentrifuge for 10 min at $4\,^{\circ}\mathrm{C}$. This lysis procedure produced intestinal extracts with protein concentrations of $\sim$1.0 to 1.5 mg/ml, as determined by the Bio-Rad Bradford method. A 70-μl aliquot of the supernatant was mixed with 10 μl of activation mix (50% glycerol, 500 μM ATP [New England BioLabs, Pickering, Ontario, Canada], 60 mM $MgCl_2$, 0.05% vol/vol Brij 35, 0.25 mg/ml bovine serum albumin [BSA]) and then incubated on the peptide array for 2 h at $37\,^{\circ}\mathrm{C}$. Arrays were then washed with PBS-1% Triton.

Slides were submerged in phospho-specific fluorescent ProQ Diamond phosphoprotein stain (Invitrogen) with agitation for 1 h before washing three times in destain solution containing 20% acetonitrile (EMD Biosciences [VWR distributor], Mississauga, Ontario, Canada) and 50 mM sodium acetate (Sigma) at pH 4.0 for 10 min. A final wash with distilled deionized H_2O was done before arrays were air dried for 20 min and centrifuged at $300 \times g$ for 2 min to remove any remaining moisture from the array. Arrays were read using a GenePix Professional 4200A microarray scanner (MDS Analytical Technologies, Toronto, Ontario, Canada) at 532 to 560 nm with a 580-nm filter to detect dye fluorescence. Images were collected using the GenePix 6.0 software (MDS). Spot intensity signals were collected as the mean of pixel intensity using the local feature background intensity calculation with the default scanner saturation level.

10.3.7 Kinome data analysis

The extent of phosphorylation of each peptide was determined as previously described [Li et al., 2012]. Briefly, local background intensities were subtracted from foreground intensities, and the resulting measurements were transformed using the variance-stabilizing normalization (VSN) [Huber et al., 2002] method to bring all the arrays onto the same scale and to eliminate variance-versus-mean dependence. The resulting data

set contained the transformed signal intensities associated with each of 300 peptides for the lysates from the different compartments within each animal, excluding the distal infected compartment from animal 1 and the proximal infected compartment from animal 4, which did not yield readable array results. Each array contained three intra-array (technical) replicates for each peptide.

As described previously [Li et al., 2012], a χ^2-test was used to identify peptides that exhibited significant technical variability, which were then excluded from subsequent analyses. Because a given animal had both treatment and control samples taken from it, the signal intensities for each infected intestinal compartment were subtracted from those of the corresponding uninfected compartment in the same animal. These will be referred to as "biological subtractions". The resulting values are presented as relative changes in kinase activity within the same animal. This approach minimizes interanimal variability, which can be significant in outbred cattle, and facilitates comparison of kinome responses to infection among animals. Hierarchical clustering was used to group the samples according to the similarity of their kinome profiles. Euclidean distance was used as the distance metric, while complete linkage was used as the linkage method. A heat map was generated wherein the columns represent samples, the rows represent peptides, and the color of each cell represents the relative level of phosphorylation for a specific peptide in a specific sample. A dendrogram representing the hierarchical clustering results for the samples is shown above the heat map. The heat maps were generated using the R function heatmap.2 from the gplots package.

Using the R function prcomp, principal-component analysis (PCA) was used to reduce the dimensionality of the kinome data. Specifically, the first three principal components (PC1, PC2, and PC3), which explain the greatest amount of variation in the kinome data, were determined. The value of each of these variables was determined for each control-subtracted compartment, and the correlations between each variable and two measures of immune response were determined as described below. The Euclidean distance was also calculated between each pair of biological subtractions. Specifically, let A_i represent the signal intensity of some treatment (e.g., distal ileum) minus the signal intensity of the control for peptide i in the same animal, and let B_i represent a different treatment minus intra-animal control subtraction. The Euclidean distance between these two biological subtractions is then $\sqrt{\sum_{i=1}^{300}(A_i - B_i)^2}$. The relationship between host kinome responses to *M. avium* subsp. *paratuberculosis* infection and immune responses of cells draining the site of infection was investigated by plotting principal components (PC1, PC2, and PC3) derived from intra-animal control-subtracted kinome profiles as a function of cellular responses to *M. avium* subsp. *paratuberculosis* lysate (IFN-γ secretion in pg/ml) and proliferation as measured by stimulation index (SI). Linear regression analysis was performed with Prism 5 for Mac OSX version 5.0b with default parameters.

10.3.8 Analysis of differentially phosphorylated peptides

A given peptide was selected for further analysis if two conditions were true: first, the peptide had to be consistently phosphorylated according to the χ^2-test for both the treatment and the control conditions; second, the P-value resulting from a t-test between the transformed treatment intensities and the transformed

control intensities had to be less than 0.2. While 0.2 may seem like a liberal threshold, when doing pathway analysis, it is more important to avoid false negatives than to avoid false positives. This is due to the fact that several peptides are involved in the same biological pathway, and we are trying to identify as many of those peptides as possible. Even with a liberal P-value threshold, it is unlikely that several peptides from the same biological pathway will be erroneously identified as differentially phosphorylated when that pathway is really not affected by the treatment under investigation; however, it increases the chances that peptides from pathways that really are affected by the treatment will be identified as differentially phosphorylated. An analysis of the impact of different P-value thresholds on false-negative probabilities can be found in the supplemental text (Appendix F).

For peptides meeting the above two conditions, fold change (FC) values were calculated using the formula 2^d, where $d = \text{average}_{\text{treatment}} - \text{average}_{\text{control}}$, as previously described [Li et al., 2012]. For visual interpretation, these peptides were input with their FC values into Ingenuity Pathway Analysis (IPA) (Ingenuity Systems, Redwood City, CA) to generate top canonical pathways with color-coded measures of relative FC values. Figures for canonical pathways were generated in IPA. Peptide lists were also uploaded to InnateDB (`http://www.innatedb.ca`), a publically available analysis resource that predicts biological pathways overrepresented in a data set and assigns a probability value (P) based on the number of proteins present for a particular pathway as well as the degree to which they are differentially expressed or modified relative to a control condition. Pathway analysis was performed in InnateDB using different FC cutoffs (1.0, 1.5, and 2.0), and pathways overrepresented for each cutoff were compiled into lists for each control-subtracted compartment. To identify common pathways overrepresented in infected compartments from the same animal and to compare with infected compartments from other animals, Venn analysis was performed with the pathway names. Common pathways of interest were further investigated by comparing individual players by Venn analysis. To uncover data trends not apparent through pathway overrepresentation analysis, gene ontology analysis (also in InnateDB) was performed by testing different FC cutoffs (1.0, 1.5, and 2.0), compiling lists, and finding common ontologies for responders and nonresponders using Venn analysis. Individual peptides that were significantly differentially phosphorylated in relation to intra-animal controls for all 6 infected compartments were also identified by Venn analysis.

10.4 Results

10.4.1 *M. avium* subsp. *paratuberculosis* infection of ileal compartments

We previously reported that surgically isolated intestinal segments prepared in 2-week-old calves could be stably maintained for up to 11 months [Charavaryamath et al., 2011], and *M. avium* subsp. *paratuberculosis* infection remained localized to individual compartments for > 9 months [Charavaryamath et al., 2013]. Here, isolated segments were divided into three compartments, and the two most distal compartments were each infected with 1×10^8 to 3×10^8 CFU of *M. avium* subsp. *paratuberculosis*, while the most proximal compart-

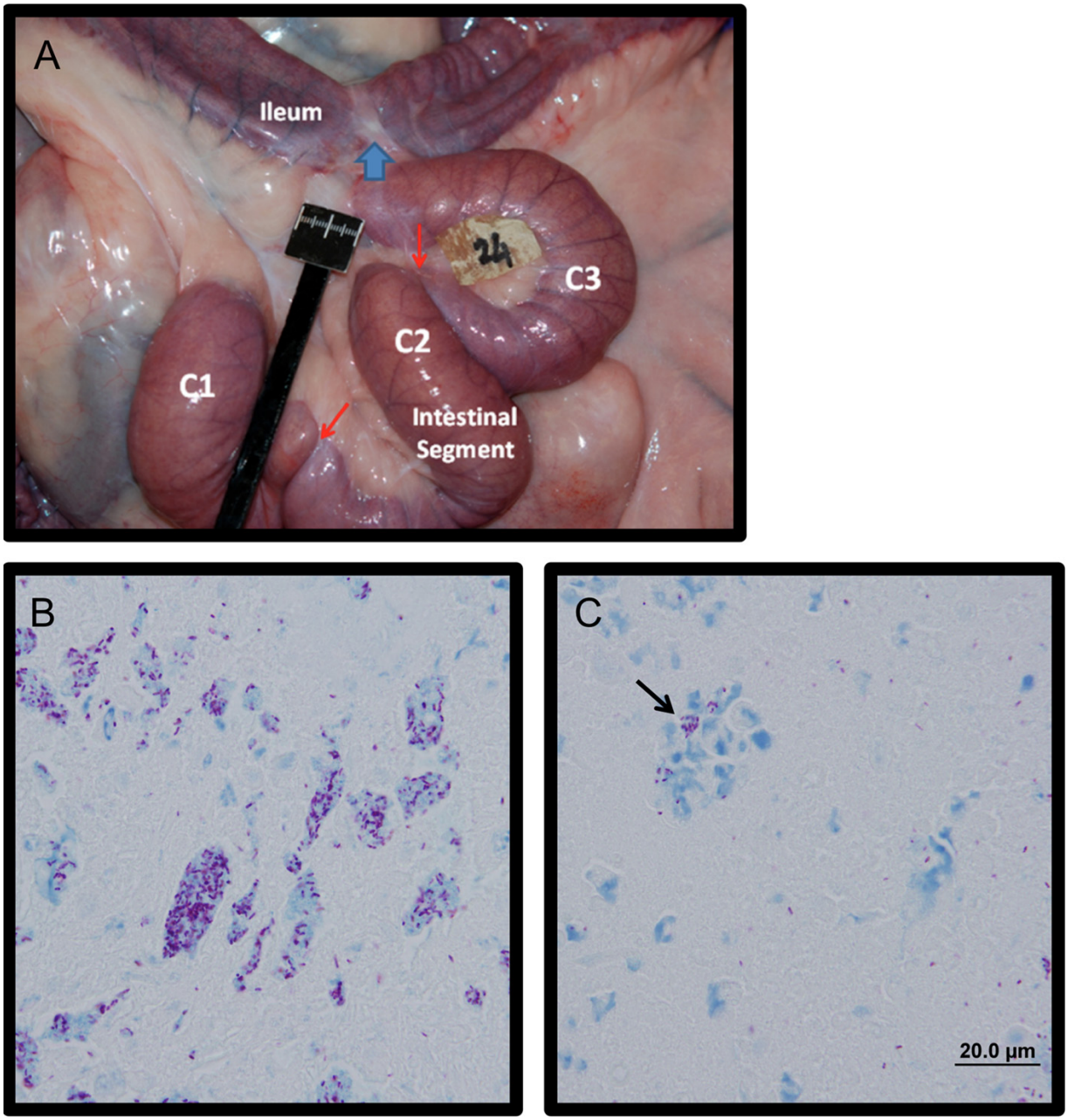


Figure 10.1: Bovine calf intestines at 1 month after *in vivo M. avium* subsp. *paratuberculosis* infection. (A) Gross appearance of a surgically isolated distal intestinal segment *in situ* at 1 month postinfection. The surgically isolated segment was subdivided into three compartments, C1, C2, and C3, using silk ligatures (indicated by red arrows). The site where intestine proximal and distal to the isolated segment was anastomosed together is indicated with a blue arrow. (B) Ziehl-Neelsen stain of intestinal contents at 1 month after infection, showing diffuse aggregates of acid-fast *Mycobacterium avium* subsp. *paratuberculosis* at a magnification of 100×. (C) Acid fast-bacteria observed within cell remnants within the intestinal contents of an infected compartment (black arrow).

ment was injected with PBS and maintained as an uninfected control. The uninfected compartment provided an intra-animal tissue reference for comparing responses to *M. avium* subsp. *paratuberculosis* challenge while controlling for changes that occur when the intestine is surgically isolated [Charavaryamath et al., 2011]. This model also facilitated isolation of cells from the specific mesenteric lymph node (MLN) draining the site of infection for assaying immune responses to *M. avium* subsp. *paratuberculosis* antigens. A clinical veterinarian observed the animals daily throughout the course of infection, and no significant changes in body temperature, weight, feed intake, or consistency of feces were noted. Gross examination of intestinal segments at the time of collection revealed no gross abnormalities (Figure 10.1A). Acid-fast bacilli were frequently observed in the intestinal contents of each infected compartment but not in those of uninfected compartments (Figure 10.1B). Furthermore, acid-fast bacilli were also frequently observed within cell remnants present within the intestinal lumen (Figure 10.1C, arrow) but were not detected within the mucosa or submucosa of the intestine. The lack of acid-fast bacilli in the intact intestinal tissue at 1 month postinfection mirrored our observation at 9 months postinfection [Charavaryamath et al., 2013]. We have previously observed early infiltration of tissues at 1, 3, and 5 days after infection under the same conditions (unpublished observations), but this early infiltration does not necessarily lead to large numbers of detectable acid-fast bacilli in the mucosa at later time points. Still, intestinal tissues within infected compartments were positive for *M. avium* subsp. *paratuberculosis* after 9 months of infection as determined by PCR [Charavaryamath et al., 2013], indicating that the host-pathogen interaction is maintained in this infection model for at least 9 months.

10.4.2 Immune responses to M. avium subsp. paratuberculosis infection

Cells isolated from the blood (PBMCs), spleen, and mesenteric lymph node (MLN) (draining the site of infection) from each calf were incubated with 1 μg/ml *M. avium* subsp. *paratuberculosis* lysate, and lymphocyte proliferation (Figure 10.2A) and IFN-γ secretion (Figures 10.2B and C) were measured. In addition to the young age of the calves when infected, the use of MLN cells specifically draining the site of infection allowed greater confidence in the specificity of the responses observed, as any environmental bacteria in the ingesta encountered during the month by the calves would not be sampled by the lymph node draining the isolated segment [Charavaryamath et al., 2013]. PBMCs from animals 3 and 4 responded with significant proliferation responses to lysate (SIs of 5.6 and 7.0, respectively), while PBMCs from animals 1 and 2 did not show significant responses. Splenocytes from all infected animals failed to proliferate in response to lysate, but MLN cells from animals 1 and 2 displayed strong proliferative responses. In contrast, MLN cells from animal 3 and 4 displayed weak proliferative responses. *M. avium* subsp. *paratuberculosis* lysate-specific IFN-γ secretion from MLN cells mirrored the results observed for proliferation, where cells from animals 1 and 2 secreted high levels of IFN-γ, while MLN cells from animals 3 and 4 secreted much lower levels (Figure 10.2B). Only PBMCs from animal 4 showed significant IFN-γ secretion (Figure 10.2C), consistent with the proliferation observed for PBMCs from the same animal (Figure 10.2A). We have observed differential immune responses similar to those seen here in 4 other calves at the 1-month time point after intestinal infection of the jejunum

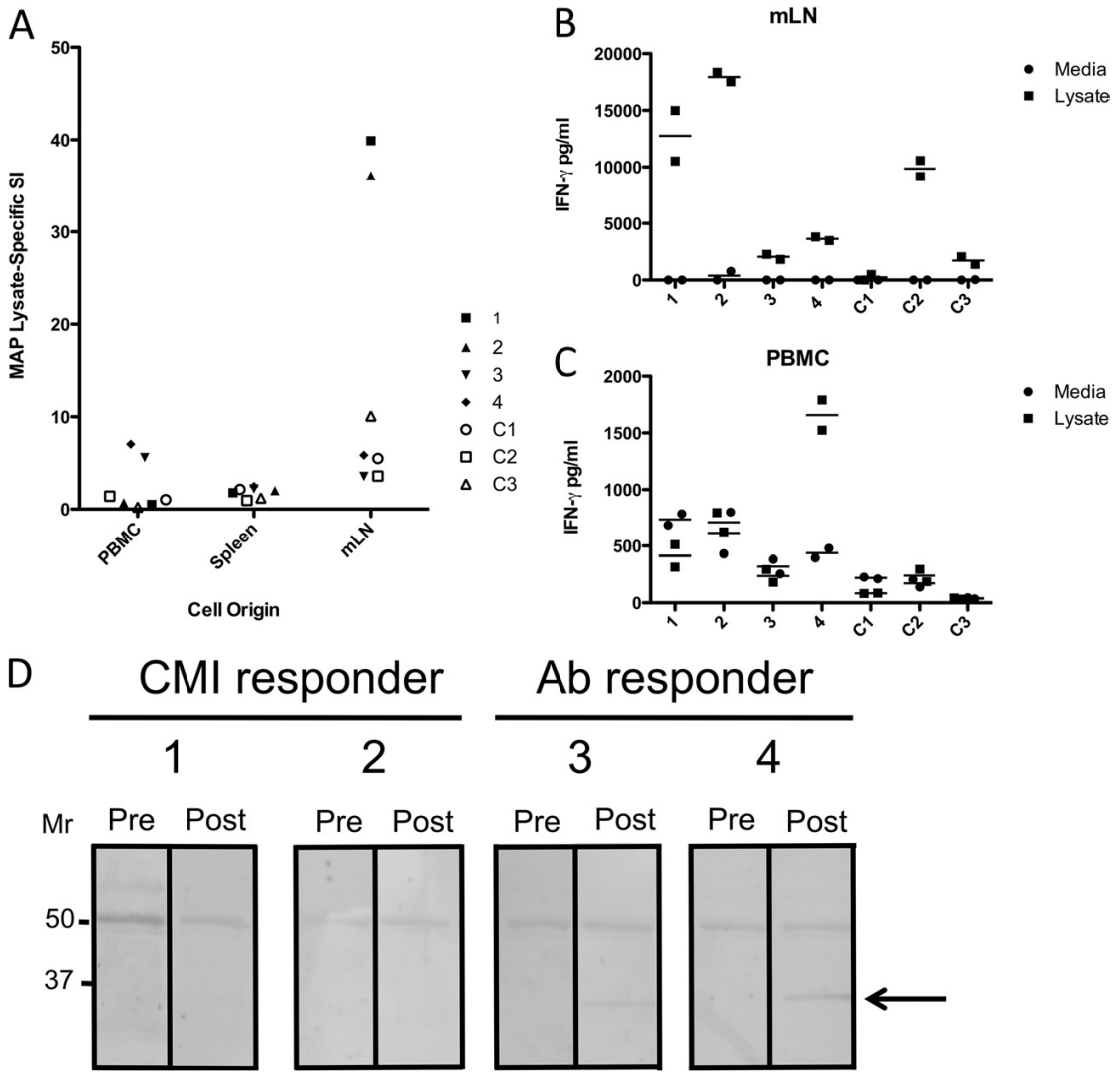


Figure 10.2: Cell-mediated and antibody immune responses of *M. avium* subsp. *paratuberculosis*-infected calves to *M. avium* subsp. *paratuberculosis* lysates. (A) Stimulation index (SI) versus cell origin for peripheral blood mononuclear cells, spleens, and mesenteric lymph nodes of four *M. avium* subsp. *paratuberculosis*-infected calves (1, 2, 3, and 4) compared to three uninfected control calves (C1, C2, and C3). (B and C) IFN-γ (pg/ml) secreted by MLN cells (B) and PBMCs (C) in response to medium or total *M. avium* subsp. *paratuberculosis* lysate from the same calves (separately stimulated duplicate well values and means are shown). (D) Immunoblots of serum collected prior to (Pre) and 1 month after (Post) experimental *M. avium* subsp. *paratuberculosis* infection against total *M. avium* subsp. *paratuberculosis* lysates. A protein of ~35 kDa detected with postinfection sera from responder calves 3 and 4 is indicated with an arrow.

Table 10.1: Euclidean distances between normalized intensity values for peptides represented on the kinome arrays. Prox, proximal compartment; Dist, distal compartment. Euclidean distances between control-subtracted compartments were calculated as described in Materials and Methods.

Compartment 1	Compartment 2	Euclidean distance
2 Prox	2 Dist	11.43
1 Prox	2 Dist	14.10
3 Prox	3 Dist	14.42
1 Prox	2 Prox	15.23
2 Dist	4 Dist	15.99
1 Prox	3 Dist	16.54
3 Prox	3 Dist	16.96
2 Prox	4 Dist	17.05
2 Dist	3 Dist	17.78
3 Dist	4 Dist	18.61
3 Prox	4 Dist	18.71
1 Prox	4 Dist	19.59
2 Dist	3 Prox	21.16
2 Prox	3 Prox	21.25
1 Prox	3 Prox	23.03

(unpublished observations), increasing our confidence that distinct immune responses can be recapitulated in the model used.

To determine if an antibody response to *M. avium* subsp. *paratuberculosis* proteins could be detected, we performed immunoblots against lysates using serum samples collected before and 1 month after infection (Figure 10.2). Both pre- and postinfection sera reacted with a band near 50 kDa, but sera from two animals reacted with an ∼35-kDa band after infection (animals 3 and 4). PBMCs isolated from the same two animals (3 and 4) showed significant proliferation and IFN-γ responses to *M. avium* subsp. *paratuberculosis* lysate (Figure 10.2A). Calves that lacked antibodies reactive to the 35-kDa protein at 1 month after infection (animals 1 and 2) (Figure 10.2D) showed strong proliferation and IFN-γ responses by MLN cells (Figure 10.2A) but not in PBMCs. Therefore, a dichotomy in *M. avium* subsp. *paratuberculosis*-specific immune responses was observed when comparing mucosal and systemic responses.

10.4.3 Kinome analysis of *M. avium* subsp. *paratuberculosis*-infected ileum

Tissue samples collected from infected and uninfected compartments were lysed, and the lysates were applied to bovine-specific kinome arrays designed as previously described [Jalal et al., 2009] and processed using

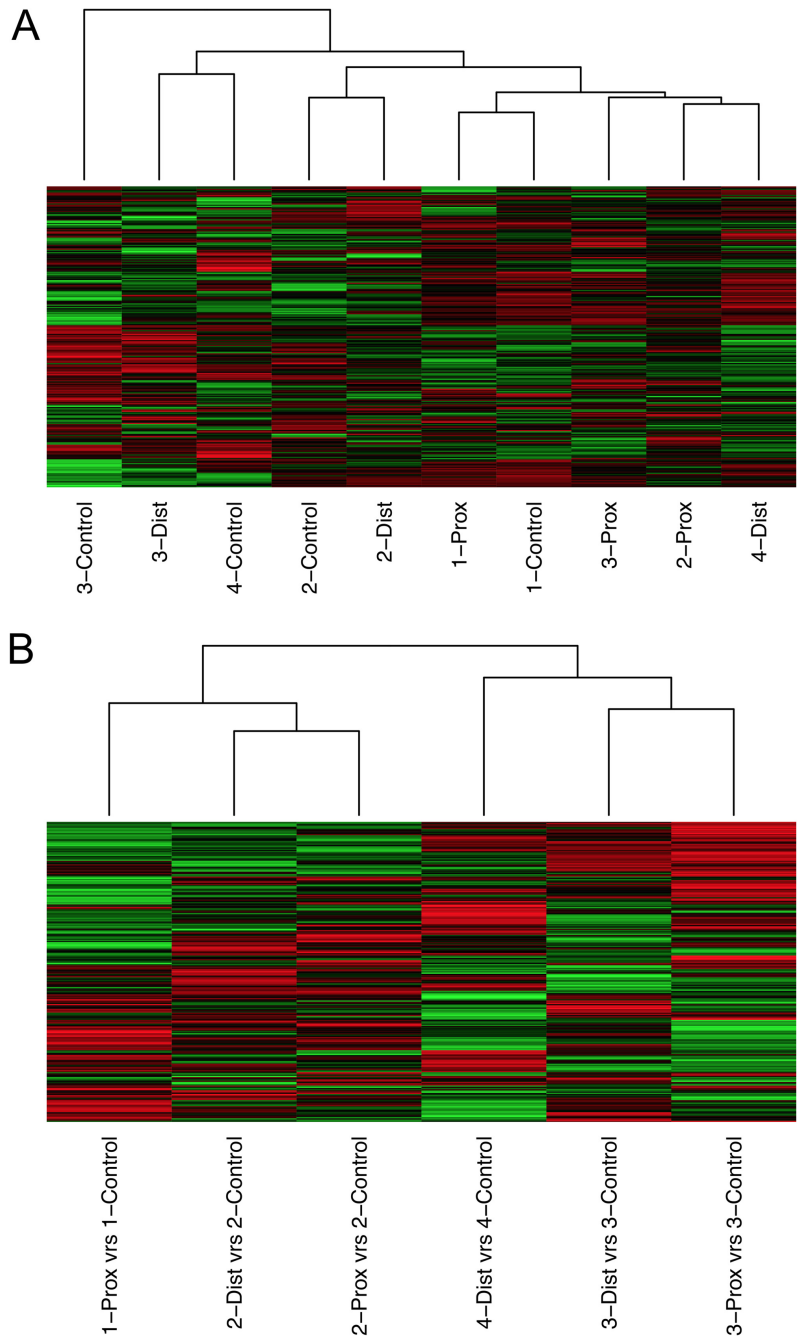


Figure 10.3: Kinome analysis of *M. avium* subsp. *paratuberculosis*-infected ileal compartments in calves. (A) Hierarchical clustering analysis to ascertain relationships between kinome responses observed in all intestinal compartments analyzed. (B) Hierarchical clustering analysis of the same responses in panel A after subtraction from responses observed for noninfected compartments in the same animal.

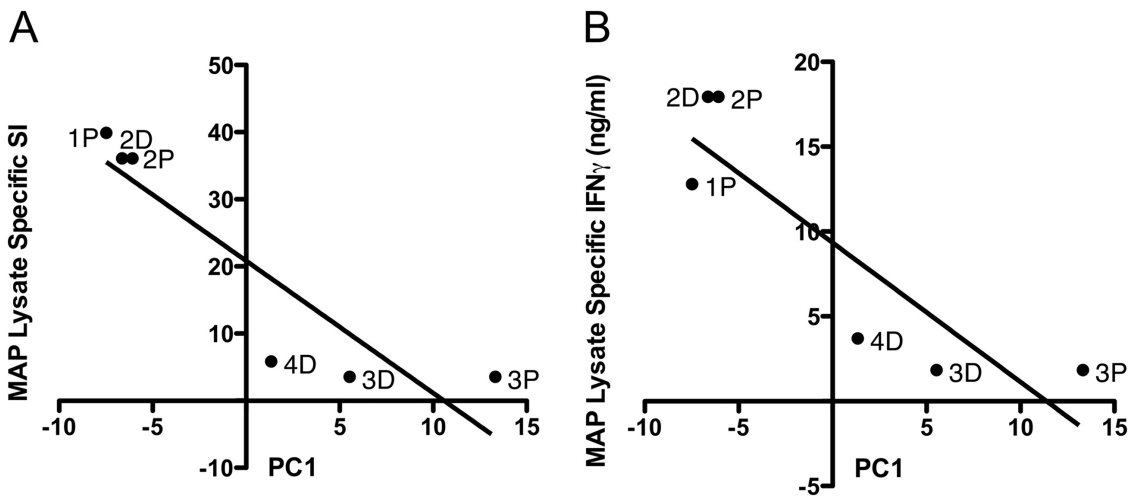


Figure 10.4: Relationships of kinome variability to cell-mediated immune responses. (A) Principal component 1 (PC1) of kinome variability for each compartment in each animal versus stimulation index ($P = 0.014$; $r^2 = 0.81$) (P, PC1 for proximal infected compartment; D, PC1 for distal infected compartment. (B) PC1 versus IFN-γ secretion (ng/ml) ($P = 0.023$; $r^2 = 0.77$).

established methods [Li et al., 2012]. Three separate compartments (2 infected and 1 uninfected) were analyzed from each animal ($n = 4$), and of the 12 compartments assayed, 10 samples provided readable results. The failed samples were from infected compartments of animals 1 and 4 (distal and proximal, respectively). For each of the treatment-control combinations, the majority of the 300 peptides (cardinality range of 280 to 294) exhibited consistent levels of phosphorylation among the technical replicates on the same array for both the treatment array and the control array. Of these peptides, those with a t-test P-value of less than 0.2 (cardinality range of 112 to 184) were chosen for subsequent analysis to focus on the most significantly altered peptides while retaining many of the target sequences.

10.4.4 Hierarchical clustering and distance calculations

The processed kinome data were subjected to hierarchical clustering using Euclidean distance as the distance metric and complete linkage as the linkage method. Kinome data sets for each intestinal compartment clustered without a clear pattern prior to subtraction of biological controls (Figure 10.3A). We have previously observed that individual animals have distinct basal kinase activities and that these distinctions must be accounted for before comparing treatments across animals [Arsenault et al., 2009, 2012, 2013a]. By considering the response of the treated condition relative to that of the control in the same animal, it was possible to determine and compare responses across animals. When intensity values for uninfected control compartments were subtracted from values for the infected compartments, the resulting data sets clustered perfectly by animal (Figure 10.3B). Across the animals, there appeared to be two separate clusters (Figure 10.3B). Animals 1 and 2 had the smallest Euclidean distance between the kinome profiles of their control-subtracted

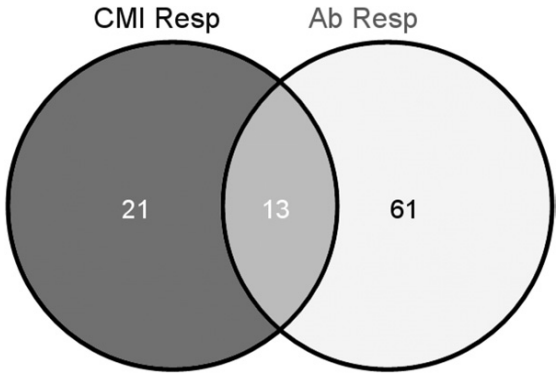


Figure 10.5: Venn analysis of significantly phosphorylated or dephosphorylated peptides shared between cell-mediated immune responder calves (CMI Resp) and antibody responder calves (Ab Resp).

compartments, while the distance between those of animals 1 and 3 was the largest (Table 10.1). These results indicate that animals 1 and 2 shared similar kinomic responses to infection, which were in turn distinct from those shared by animals 3 and 4 (Figure 10.3A).

10.4.5 Linear regression analysis of kinome profiles versus cellular responses to *M. avium* subsp. *paratuberculosis* lysates

We observed that the animals that clustered together in the hierarchical clustering analysis after subtraction of uninfected control responses (Figure 10.3B) also showed similar *M. avium* subsp. *paratuberculosis* lysate-specific proliferation, IFN-γ, and antibody responses (Figure 10.2). To further examine this relationship, PCA was carried out, allowing us to represent much of the variability in the kinome data by three variables, denoted PC1, PC2, and PC3. PC1 captured a large proportion of the variability (approximately 31%), and the value of PC1 for each control-subtracted compartment was plotted as a function of *M. avium* subsp. *paratuberculosis*-lysate specific proliferation (Figure 10.4A) and IFN-γ secretion (Figure 10.4B) of MLN-derived cells. Linear regression analysis confirmed a significant negative linear correlation between PC1 and both *M. avium* subsp. *paratuberculosis* lysate-specific SI ($P = 0.014$) (Figure 10.4A) and IFN-γ secretion ($P = 0.023$) (Figure 10.4B). Neither PC2 nor PC3 showed a significant linear relationship with the MLN cell responses. These results revealed that significant differences in the overall intestinal kinome profiles of infected compartments from animals 1 and 2 or animals 3 and 4 could be correlated with the distinct responses of cells from these animals to *M. avium* subsp. *paratuberculosis* lysates.

10.4.6 Analysis of kinome array data

As mentioned above, hierarchical clustering of the kinome data, as well as the Euclidean distance calculations, indicated that the infected compartments from the same animals clustered closest together, followed by a

Table 10.2: Select CREB and Wnt/β-catenin pathway peptide phosphorylation sites differentially phosphorylated by *M. avium* subsp. *paratuberculosis*-infected intestinal lysates from CMI responder and antibody responder calves. The fold change (FC) and probability value from a t-test between the transformed treatment (infected lysate) intensities and the transformed control (uninfected lysate) intensities for the indicated peptide (Uniprot accession numbers are given in parentheses) and residue(s) are shown. Prox, proximal compartment; Dist, distal compartment. —, inconsistently phosphorylated.

	CMI responders				Antibody responders					
	LEF1 (Q9UJU2); T155 or S166 (↑)		β-Catenin (P53222); Y142, T41/45, or S675 (↑)		ADCY8 (P40145); Y406 (↑)		p300 (Q09472); S893 or S1834 (↑)		PP2CA (P67775); T304/7 (↑)	
	FC	P	FC	P	FC	P	FC	P	FC	P
1 Prox	1.5	0.08	1.7	0.13	—	—	−1.6	0.10	1.2	0.31
2 Prox	5.5	0.04	1.6	0.06	−1.4	0.36	−1.6	0.23	−1.4	0.1
2 Dist	3.9	0.07	1.7	0.18	−1.0	0.48	−1.5	0.04	−1.5	0.3
3 Prox	−2.0	0.03	−3.4	0.0004	16	0.02	6.2	0.005	1.5	0.04
3 Dist	−1.8	0.26	−3.2	0.02	6.4	0.04	3.4	0.04	2.3	0.12
4 Dist	−2.2	0.05	−1.8	0.006	6.5	0.01	2.0	0.1	1.2	0.14

close grouping of animals 1 and 2 (cell-mediated immunity [CMI] responders) distinct from animals 3 and 4 (antibody responders) (Figure 10.3B; Table 10.1). Peptides that were significantly phosphorylated or dephosphorylated relative to the uninfected control ($P < 0.2$) were compiled for the different groupings (CMI responders versus antibody responders) and compared by Venn analysis (Figure 10.5). The greatest similarity was observed between antibody responders (animals 3 and 4), as predicted from the clustering (61 shared significantly altered peptides). Of the 13 peptides that were consistently differentially phosphorylated across animals, 4 belong to a pathway called "inactivation of gsk3 by AKT causes accumulation of β-catenin" (Pathway Interaction Database [PID] BioCarta 4022). To observe trends in the peptides that differed between the groups, data sets ($P < 0.2$) were uploaded to Ingenuity pathway analysis (IPA) for visual comparisons. IPA determines the top canonical pathways represented within each data set, and it identified "molecular mechanisms of cancer" as the top canonical pathway for all 6 infected compartments analyzed. This allowed for an overall visual comparison of the top pathway for all the data sets in terms of increased or decreased phosphorylation (Figure 10.6A and B). It was immediately apparent that animals 1 and 2 showed distinctly higher phosphorylation of many players within this set of pathways than animals 3 and 4, where decreases in phosphorylation predominated for the same targets. Peptides corresponding to the transcription factor LEF-1 and several other players in the Wnt/β-catenin pathway were highly phosphorylated by *M. avium* subsp. *paratuberculosis*-infected ileal lysates from animals 1 and 2 relative to uninfected ileal lysates from the same animal. Conversely, the same targets were less phosphorylated by infected intestinal lysates from animals 3 and 4 relative to their intra-animal controls (Figure 10.6, right panel, and Table 10.2). The opposite was true for some of the peptides representing proteins involved in the Wnt/β-catenin pathway, such as ADCY8 and p300 (Table 10.2).

Pathway overrepresentation analysis and gene ontology overrepresentation analysis are often used to focus

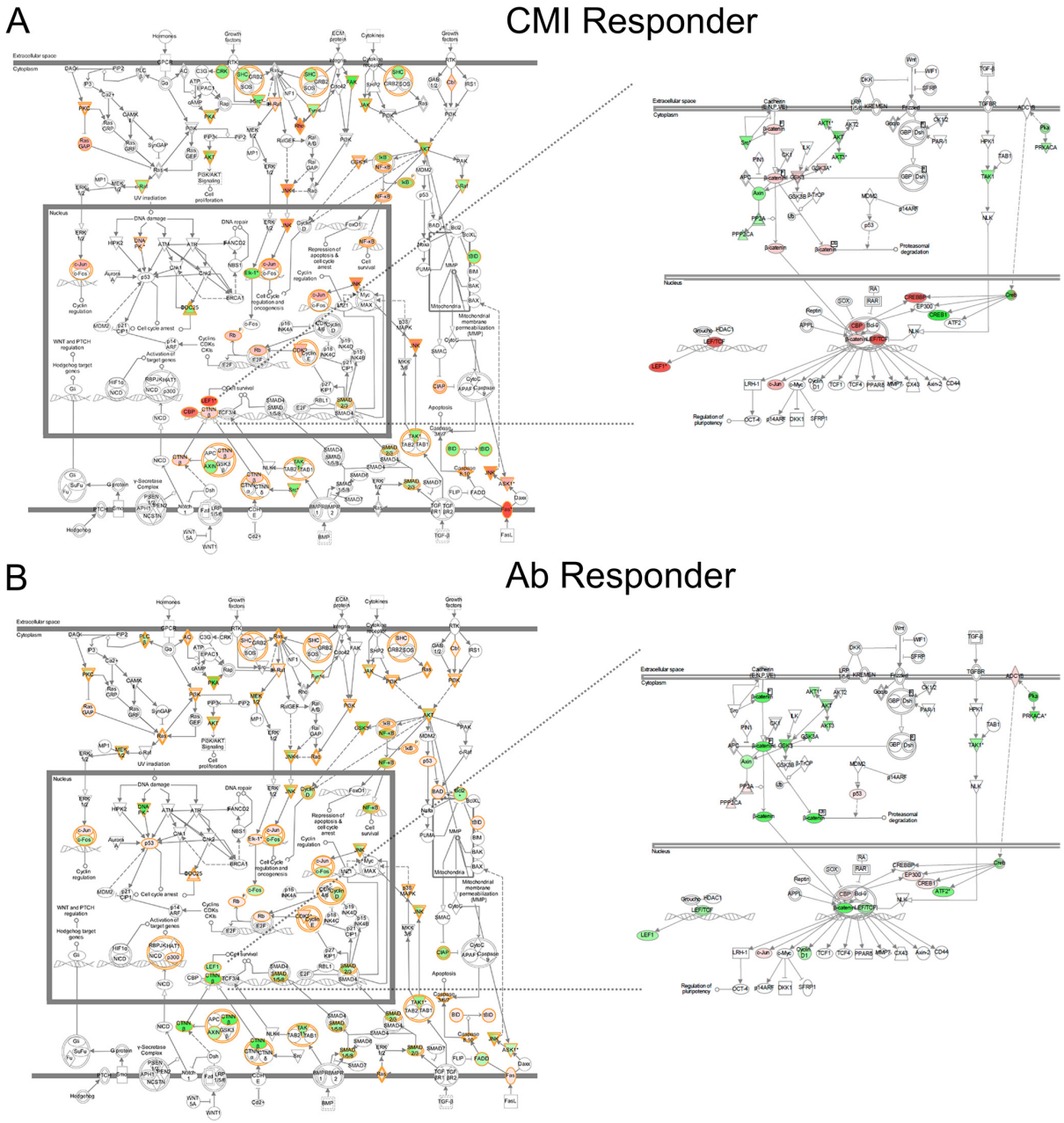


Figure 10.6: Ingenuity pathway analysis (IPA) of kinome profiles, showing top canonical pathway differences between cell-mediated immune responder (CMI Responder) and antibody responder (Ab Responder) calves. The intensity of the color depicts the relative increase (red) or decrease (green) in phosphorylation. (A) Left panel, CMI responder top pathway. Right panel, zoomed-in view of Wnt/β-catenin pathway for the same analysis. (B) Left panel, Ab responder top pathway. Right panel, zoomed-in view of Wnt/β-catenin pathway for the same analysis. Analyses shown are for the animal 2 proximal *M. avium* subsp. *paratuberculosis*-infected compartment (CMI responder) and the animal 3 proximal *M. avium* subsp. *paratuberculosis*-infected compartment (Ab responder).

on specific pathways altered within a homogeneous cell population. Here we used these methods to identify trends in the global tissue profile of kinase activity against the array peptides. Significantly increased or decreased phosphorylation of multiple targets within a similar biological process or pathway provides greater confidence in the trends observed but must also be considered with caution because of the averaging of the represented kinases during whole-tissue lysis. We did not find significantly altered common pathways or gene ontologies shared across all 6 infected compartments, but we did identify several pathways and ontologies shared among infected compartments from animals with similar kinome profiles. Animals 1 and 2 exhibited increased phosphorylation of players in the interleukin-1 (IL-1) and transforming growth factor β (TGF-β) signaling through TAK1 pathways, while animals 3 and 4 showed increased phosphorylation of players in the IL-6, natural killer cell-mediated cytotoxicity, and IL-4 signaling pathways (Table 10.3). Gene ontology overrepresentation analysis revealed significantly increased phosphorylation of players in the innate immune response and decreased phosphorylation of players in peptidyl-tyrosine phosphorylation for lysates of infected compartments from animals 1 and 2 (Table 10.4). Gene ontologies over-represented in the arrays exposed to lysates from animals 3 and 4 included increased phosphorylation of players involved in epidermal growth factor (EGF) receptor signaling and decreased phosphorylation of players in the Wnt receptor signaling pathway, matching the observed decrease in phosphorylation of Wnt pathway players by visual assessment of top canonical pathways using IPA.

10.5 Discussion

We previously used kinome arrays to address how *M. avium* subsp. *paratuberculosis* modulates host signaling in isolated infected bovine monocytes [Arsenault et al., 2012, 2013a]. Here, we infected surgically isolated intestinal compartments with a controlled dose of *M. avium* subsp. *paratuberculosis* and collected intestinal tissue for analysis at 1 month postinfection. Kinome arrays were used to measure global changes in tissue kinase activity relative to intra-animal naive control intestinal compartments. The control-subtracted whole-tissue kinome profiles cluster by animal and then by the type of cell-mediated or antibody responses mounted against *M. avium* subsp. *paratuberculosis* lysate. A small number of the proteins represented on the arrays (13) were significantly differentially phosphorylated in the infected compartments compared to uninfected compartments across all 6 biological replicates. However, major differences between pathways, players, and ontologies were observed for animals that showed different immune response profiles, suggesting that global intestinal kinome profiles reflect different host responses following *M. avium* subsp. *paratuberculosis* infection.

Kinome analysis of a homogeneous cell population, such as *M. avium* subsp. *paratuberculosis*-infected monocytes, and kinome profiles of whole-tissue lysates provide distinct information regarding host responses. When studying a homogeneous cell population, the kinome arrays can provide insight into specific cell signaling pathways directly altered by *M. avium* subsp. *paratuberculosis* infection [Arsenault et al., 2012, 2013a]. In contrast, kinome data from tissue lysates reflect the average effective kinase activity within the

Table 10.3: Pathway overrepresentation analysis of CMI responder and antibody responder calves and associated probabilities of upregulation as determined by InnateDB. The number of peptides showing increased (↑) or decreased (↓) phosphorylation and associated probabilities of upregulation (↑) for the indicated pathway (database identification numbers are in parentheses) are shown. Prox, proximal compartment; Dist, distal compartment.

	CMI responders						Antibody responders								
	IL-1 (NETPATH 10429) (↑)			TGF-β (INOH 10330) (↑)			IL-6 (NETPATH 10415) (↑)			NK cell (KEGG 578) (↑)			IL-4 (NETPATH 10417) (↑)		
	↑	↓	P	↑	↓	P	↑	↓	P	↑	↓	P	↑	↓	P
1 Prox	11	4	0.06	6	0	0.004	5	5	0.66	3	3	0.55	6	7	0.53
2 Prox	10	4	0.05	3	1	0.24	7	7	0.61	4	5	0.66	7	5	0.81
2 Dist	7	3	0.26	5	1	0.08	6	7	0.37	2	6	0.92	4	4	0.47
3 Prox	9	7	0.84	3	4	0.92	22	9	0.04	12	1	0.01	13	2	0.14
3 Dist	8	7	0.89	2	4	0.94	17	8	0.12	9	5	0.30	10	4	0.11
4 Dist	9	10	0.87	4	5	0.79	19	8	0.06	13	4	0.05	15	5	0.08

diverse cell populations sampled, including both direct and indirect effects, and must be understood as a tissue signature rather than representative pathways for a single cell type. Assaying the overall kinase activities within a tissue and subtracting these values from those for an intra-animal control allows for identification of the most significantly altered kinase activities throughout the tissue. The potential exists that these averaged kinase activities will fail to detect important kinase activity occurring in relatively rare cells such as dendritic cells or macrophages that perform critical functions in defining host-pathogen interactions.

Whole-tissue kinase activity measurements in some ways resemble whole-tissue gene expression profiling, where consistent sampling is important [Mutch et al., 2009], and results must be interpreted as overall averages of expression. Tissue gene expression has been useful for comparative studies of cancers [Sanz-Pamplona et al., 2012] that can yield valuable *in vivo* insights [Grzmil et al., 2011]. To date, significant success has been achieved for tissue studies focused on kinase activity [de Borst et al., 2007], especially in studies of brain [Sikkema et al., 2009, Hoozemans et al., 2012], where high numbers of distinct cell types are represented in a given tissue sample.

The kinome profiles generated with whole intestinal samples indicated that *M. avium* subsp. *paratuberculosis* infection can result in divergent kinase activity in different calves. The distinct tissue kinase signatures of different animals could be related to differential recruitment of specific immune effector cell populations [Charavaryamath et al., 2013]. Differences in kinome profiles are also likely influenced by the unique genotypes of outbred calves, and it is likely that a broader range of kinome profiles will be observed as more animals are studied. Genotypes may confer enhanced susceptibility or resistance to *M. avium* subsp. *paratuberculosis* infection [Minozzi et al., 2012], but the exact linkages are complex and remain undefined. On the other hand, kinase activity provides a more direct measure of phenotypic responses and may provide more effective parameters to identify cattle that mount protective immune responses to infection. Kinome responses may also uncover links to help better characterize protective genotypes. It is understood that *M. avium* subsp. *paratuberculosis* subverts the host immune system by inhibiting phagosome-lysosome fu-

Table 10.4: Gene ontology analysis of CMI responders and antibody responders and associated probabilities of up- or downregulation as determined by InnateDB. The number of peptides and associated probabilities of upregulation (↑) or downregulation (↓) for the indicated ontology (database identification numbers are in parentheses) are shown. Prox, proximal compartment; Dist, distal compartment. —, ontology was not represented in the analysis.

	CMI responders						Antibody responders								
	Innate immune response (GO:0045087) (↑)			Peptidyl-tyrosine phosphorylation (GO:0018108) (↓)			Positive regulation of DNA replication (GO:0045740) (↑)			Epidermal growth factor receptor signaling (GO:0007173) (↑)			Wnt receptor signaling pathway (GO:0016055) (↓)		
	↑	↓	*P*	↑	↓	*P*	↑	↓	*P*	↑	↓	*P*	↑	↓	*P*
1 Prox	17	11	0.07	2	4	0.37	—	—	—	2	3	0.80	0	0	1
2 Prox	23	17	0.10	3	8	0.08	2	1	0.38	3	4	0.63	2	2	0.53
2 Dist	27	15	0.03	3	7	0.09	1	0	0.53	3	1	0.28	1	0	1
3 Prox	33	21	0.40	7	3	0.79	4	0	0.09	9	1	0.02	1	4	0.07
3 Dist	22	16	0.77	7	2	0.86	5	0	0.05	9	3	0.01	1	2	0.25
4 Dist	34	23	0.32	5	4	0.86	4	0	0.07	9	2	0.09	1	5	0.06

sion in macrophages that take up the bacterium [Woo et al., 2007]. Immune evasion by *M. avium* subsp. *paratuberculosis* involves a variety of other mechanisms, including evading cell-mediated immunity through enhancing secretion of suppressor cytokines, activating T-regulatory cells, inhibiting tumor necrosis factor alpha (TNF-α) expression, and inhibiting cytotoxic killing of infected cells [Coussens, 2004]. One of the antigens recognized by *M. avium* subsp. *paratuberculosis* infected cattle is a 35-kDa major membrane protein (MMP) [Bannantine et al., 2003, Shin et al., 2005] that most likely corresponds to the protein detected with sera from two of the calves in the current study (Figure 10.2). Antibodies do not appear to confer protection against progression of the primarily intracellular infection [Coussens, 2004]. While we cannot predict which early responses to *M. avium* subsp. *paratuberculosis* observed at 1 month postinfection would be the most effective for clearance of the pathogen, longer-term studies that correlate early responses with chronic infection or clearance may reveal the most effective early responses. Kinome profiling may be one way to distinguish protective versus nonprotective early responses to *M. avium* subsp. *paratuberculosis* infection.

Due to the small sample size, it is difficult to determine the significance of specific differences in kinase activities among the animals exhibiting different immune responses following *M. avium* subsp. *paratuberculosis* infection. Further studies with more animals will be required to more confidently determine the kinomic correlates of specific immune responses and more fully define variations that may be influenced by other environmental factors. However, the general increase in phosphorylation of Wnt/β-catenin, IL-1, and TGF-β (through TAK1) pathway proteins in animals that showed strong MLN proliferation and IFN-γ responses to *M. avium* subsp. *paratuberculosis* lysate compared to animals that did not show MLN responses to lysates warrants further investigation. Similarly, the increased phosphorylation of IL-6, NK cell, and IL-4 pathway players in animals that showed antibody responses but not MLN proliferation responses to lysate indicates global differences in kinase activity at the site of infection that may reflect either protective or nonprotective

responses to *M. avium* subsp. *paratuberculosis* infection.

Interestingly, a recent short-term study (0.5, 1, 2, 4, 8, and 12 h) of experimental *M. avium* subsp. *paratuberculosis* infection in ileal compartments revealed shifts in gene expression patterns in 4 calves over the course of infection [Khare et al., 2012]. The authors did not comment on differences between calves, perhaps because very-early-stage responses to infection are less variable. However, the shifts they observed included early suppression (at 0.5 and 1 h postinfection) of the Wnt receptor signaling pathway through β-catenin, followed by late-phase activation of the same pathway (at 12 h postinfection) (among several other pathways). They also observed "late-phase" activation of genes implicated in innate immune response gene ontology, which they suggested to be indicative of an effective immune response. Here, we observed increased phosphorylation of innate immune response genes by ileal lysates collected 1 month after infection, most significantly in CMI responder animals, and decreased phosphorylation of Wnt receptor signaling pathway players in animals that failed to mount local *M. avium* subsp. *paratuberculosis*-specific CMI but did mount antibody immune responses. These observations seem to suggest a divergence of kinase activity at 1 month postinfection that varied among individual animals. Further studies will be necessary to determine if specific kinase activities may correlate with effective bovine responses to *M. avium* subsp. *paratuberculosis* infection and exactly what interval after infection is required before animal-specific differences may be observed. The correlation of specific immune responses against lysates with general kinase activity at the site of infection provides a method to evaluate the effectiveness of host evasion by *M. avium* subsp. *paratuberculosis* and may uncover strategies to promote pathogen clearance.

10.6 Acknowledgments

This research was funded by the Saskatchewan Agriculture Development Fund (ADF). Philip Griebel is a holder of a Tier I CRC in Mucosal Immunology, which is funded by the Canadian Institutes of Health Research. We are grateful to the VIDO Animal Care staff for surgeries, veterinary care, animal handling, and help with sample collection. We thank Donna Dent for help with IFN-γ ELISAs, Chris Stuart from the Western College of Veterinary Medicine for help with microscopy, and Natasa Arsic for help with isolation of splenocytes. We also thank Qingxiang Yan for statistical assistance.

Chapter 11

Identification of developmentally-specific kinotypes and mechanisms of Varroa mite resistance through whole-organism, kinome analysis of honeybee

Albert J Robertson, Brett Trost, Erin Scruten, Thomas Robertson, Mohammad Mostajeran, Wayne Connor, Anthony Kusalik, Philip Griebel and Scott Napper

In Chapter 6, the application of DAPPLE to the design of a honeybee-specific kinome array was described. Chapter 6 was meant to be a case study in the use of DAPPLE to design arrays for species that are distantly related to the species represented in the phosphorylation site databases. As such, it did not describe the application of the honeybee arrays. This chapter, which presents the last of three papers that describe biological applications of the work described in this thesis, aims to fill that gap. Specifically, it describes the application of kinome arrays (as well as other biological techniques) to the study of honeybees (*Apis mellifera*) that differ in their developmental stage, susceptibility to infestation by the mite *Varroa destructor*, and infestation status (infested or uninfested by Varroa).

Citation

A. J. Robertson, B. Trost, E. Scruten, T. Robertson, M. Mostajeran, W. Connor, A. Kusalik, P. Griebel and S. Napper. Identification of developmentally-specific kinotypes and mechanisms of Varroa mite resistance through whole-organism, kinome analysis of honeybee. *Front Genet* 5:139, 2014.

Author contributions

Albert Robertson designed the breeding program, helped plan the kinome array experiments, participated in data analysis, supervised the research, and wrote parts of the manuscript. Brett Trost helped design the honeybee-specific kinome arrays, participated in data analysis, and wrote parts of the manuscript. Erin Scruten performed the kinome array experiments. Thomas Robertson and Mohammad Mostajeran performed

bee selections and helped define the resistant and susceptible phenotypes. Wayne Connor performed the virus quantification experiments. Anthony Kusalik and Philip Griebel participated in data analysis and supervised the research. Scott Napper helped design the honeybee-specific arrays, participated in data analysis, wrote parts of the manuscript, and supervised the research. All authors participated in revising the manuscript, with particular contributions by Albert Robertson, Brett Trost, Anthony Kusalik, and Scott Napper.

Supplementary material

Supplementary tables for this paper are given in Appendix G. This paper also includes one supplementary file, which can be downloaded from `http://saphire.usask.ca/saphire/honeybee/honeybee_array.gal`.

11.1 Abstract

Recent investigations associate *Varroa destructor* (Mesostigmata: Varroidae) parasitism and its associated pathogens and agricultural pesticides with negative effects on colony health, resulting in sporadic global declines in domestic honeybee (*Apis mellifera*) populations. These events have motivated efforts to develop research tools that can offer insight into the causes of declining bee health as well as identify biomarkers to guide breeding programs. Here we report the development of a bee-specific peptide array for characterizing global cellular kinase activity in whole bee extracts. The arrays reveal distinct, developmentally-specific signaling profiles between bees with differential susceptibility to infestation by Varroa mites. Gene ontology analysis of the differentially phosphorylated peptides indicates that the differential susceptibility to Varroa mite infestation does not reflect compromised immunity; rather, there is evidence for mite-mediated immune suppression within the susceptible phenotype that may reduce the ability of these bees to counter secondary viral infections. This hypothesis is supported by the demonstration of more diverse viral infections in mite-infested, susceptible adult bees. The bee-specific peptide arrays are an effective tool for understanding the molecular basis of this complex phenotype as well as for the discovery and utilization of phosphorylation biomarkers for breeding programs.

11.2 Introduction

In recent years, there has been an alarming worldwide decline in populations of honeybees (*Apis mellifera*) [Dietemann et al., 2013]. This is of considerable concern, as approximately one-third of the human food supply depends on pollination by the honeybee [Greenleaf and Kremen, 2006, Cox-Foster et al., 2007, Vanengelsdorp et al., 2009]. A number of possible causes have been suggested, including Varroa mite parasitism and associated pathogens [Nazzi et al., 2012, Martin et al., 2012], increased use of pesticides, lack of genetic diversity, and other factors [Vanengelsdorp et al., 2009, Mullin et al., 2010].

The ectoparasitic mite *Varroa destructor*, and RNA viruses that are associated with it, are a significant challenge to the honeybee. Deformed wing virus (DWV) [Martin et al., 2012, 2013], Israeli acute paralysis virus (IAPV), acute bee paralysis virus (ABPV), and Kashmir bee virus (KBV) are the major viruses vectored by Varroa [Di Prisco et al., 2011]. Varroa mites continue to spread throughout the world and contribute to the decline of domesticated honeybee populations [Nazzi et al., 2012, Martin et al., 2012]. Their natural host, the Asian honeybee (*Apis ceranae*), has developed protective mechanisms based on behavioural characteristics, such as grooming and hygienic traits, as well as differences in brood development time, rather than differences in immunity [Sammataro et al., 2000, Rosenkranz et al., 2010]. The western honeybee, initially exposed to Varroa mite parasitism in the mid-1960s [Sammataro et al., 2000], has yet to develop adequate resistance mechanisms. Many synthetic miticides have been deployed to combat Varroa infestations, but the mites quickly develop resistance; further, the miticides have detrimental effects on honeybee health, and can also

leave dangerous residues in the wax [Lodesani and Costa, 2005].

A more attractive approach is to breed honeybees capable of resisting or controlling Varroa mite infestation. However, breeding for Varroa resistance is complicated by a lack of understanding of honeybee susceptibility to mite parasitism, a dearth of biomarkers to identify potentially resistant progeny, and the instability of resistant phenotypes. A number of groups have used natural selection to identify colony phenotypes with Varroa resistance [Le Conte et al., 2007, Seeley, 2007]. The most well-characterized genetic stocks able to suppress Varroa population growth are the Varroa sensitive hygiene (VSH) lines [Harbo and Harris, 2009, Tsuruda et al., 2012]. In this work, the Saskatraz natural selection project (`http://www.saskatraz.com`) selected and characterized susceptible and resistant honeybee colony phenotypes for molecular analyses. This project focuses on recurrent natural selection of survivor colonies for honey production, wintering ability, resistance to Varroa, and overall colony health, in the absence of synthetic miticides.

There is a general consensus that understanding the cellular mechanisms of these disease-resistance phenotypes requires a global perspective on bee biology. To this end, a number of recent studies have examined the differential expression of genes [Le Conte et al., 2011] and proteins [Parker et al., 2012] in honeybees that suppress Varroa population growth. These efforts have neither provided clear insight into the cellular mechanisms of Varroa mite susceptibility nor identified reliable biomarkers. This reflects the challenges associated with deciphering complex biology, in particular within the context of a mixed genetic population.

Similar challenges have been overcome in other livestock species through the development and application of species-specific peptide arrays for analysis of global cellular kinase (kinome) activity [Arsenault et al., 2012, Trost et al., 2013a, Arsenault et al., 2013b]. Kinase-mediated protein phosphorylation is critical for the regulation of cellular responses and phenotypes. Analysis of global kinome activity has provided a powerful tool to understand complex biology as well as to identify therapeutic targets and biomarkers [Eglen and Reisine, 2011]. In particular, the ability to use short peptides as surrogate substrates for kinases makes it possible to monitor the kinome using high-throughput peptide arrays [Arsenault et al., 2011]. While detailed descriptions of the phosphoproteome are available for only a limited number of species, it is possible to predict the sequence contexts of phosphorylation events based on genomic information, creating the opportunity to develop species-specific kinome microarrays for species whose phosphoproteomes have not been extensively characterized [Jalal et al., 2009, Trost et al., 2013a]. Kinome analysis has been demonstrated to have considerable utility in understanding cellular mechanisms of host-pathogen interaction [Kindrachuk et al., 2011, Arsenault et al., 2012, 2013a, Määttänen et al., 2013, Mulongo et al., 2014] as well as identifying phosphorylation biomarkers that predict or reflect phenotypic traits [Arsenault et al., 2013b]. Recently, the existence of temporally-stable species and individual-specific phosphorylation profiles, or kinotypes, was reported [Trost et al., 2013c]. These stable patterns within individuals likely reflect genetic, epigenetic, environmental and developmental influences and may provide mechanistic and predictive insight into complex, multi-factorial phenotypes. Similarly, while kinome analysis is traditionally performed on samples of low biological complexity, such as cultured cells or purified cell populations, recent applications have extended this analysis to

more complex samples, including intestinal tissue [Määttänen et al., 2013] and muscle biopsies [Arsenault et al., 2013b].

Here we report the development of a bee-specific kinome array and its application to characterize honeybees with a quantified, differential susceptibility to Varroa mite infestation. Bees of the susceptible and resistant phenotypes possess distinct kinome profiles at a number of developmental stages ranging from pupae to adult, highlighting the potential to use these differences as markers for breeding programs. Kinome analysis also offers insight into the mechanisms underlying disease susceptibility. Specifically, the kinome data indicate that the susceptibility to Varroa mite infestation does not reflect compromised immunity. There is, however, evidence for mite-mediated immune suppression within the susceptible phenotype, which may reduce the ability of these bees to counter secondary infections. Consistent with this hypothesis, an increased diversity of viral infections is observed in Varroa-infested susceptible bees. Overall, the bee-specific peptide arrays offer an effective tool for understanding the molecular basis of complex phenotypes and for analyzing specific biological responses, and may facilitate the identification of phosphorylation biomarkers for breeding programs.

11.3 Materials and methods

11.3.1 Colony phenotype selection

A detailed description of the honeybee breeding and selection program that was used to construct and identify the Varroa mite susceptible and resistant phenotypes can be accessed at `http://www.saskatraz.com`. Briefly, Meadow Ridge Enterprises Ltd. established a closed-population mating program in 1992, selecting from approximately 1,200 colonies annually for honey production, wintering ability and chalk brood resistance. Tracheal mites were first observed in the colonies in the late 1990s, and Varroa mites were detected shortly thereafter. The selected population showed no resistance to either mite. To introduce mite resistance, Russian stock was imported as embryos from the USDA between 2001 and 2005 [Rinderer et al., 2001]. Russian virgins from three different selections were close-population mated to selected colonies at the Meadow Ridge apiary. The F1 hybrids from these initial crosses were established at three different isolated apiaries, and used to backcross Russian virgins from subsequent shipments to regenerate Russian stock, and for re-selection under Canadian conditions. These apiaries served as a source of colonies for the natural selection apiary, and for drones in crosses used to increase Varroa resistance. In 2004, a natural selection apiary was established at an isolated area in Saskatchewan, called Saskatraz, using colonies from Meadow Ridge and collaborating Saskatchewan beekeepers. This apiary was established to further select for productive colonies with mite resistance and good wintering ability, without synthetic miticide treatment. Tracheal mites were introduced in the fall of 2004 by adding 200 worker bees with 60% tracheal mite infestations. Varroa mites were present in the original selections.

A colony phenotype called Saskatraz 88 (S88) was constructed by backcrossing a daughter from a Russian

hybrid line selected at Saskatraz in 2006 to drones at an isolated Russian apiary (RP30) previously established at Meadow Ridge to increase Varroa tolerance. The resulting colony superseded and a daughter was mated at the RP30 apiary again, resulting in two back crosses at the RP30 apiary. Extensive screening of Varroa present on adult bee populations in both breeding populations and commercial colonies identified G4, a susceptible colony phenotype established in the summer of 2009. G4 bees showing high Varroa mite infestations during spring evaluations were selected and moved to an isolated apiary used as a Varroa nursery for experimental purposes. Susceptible colonies were not treated and left to die, serving to remove susceptible colonies from the breeding population. G4 and S88 were located in different apiaries during the course of the experiment. No queen events (swarming, supersedure) were noted in either S88 or G4 colonies during their lifespans. The S88 queen was last observed in the fall of 2010 in the Saskatraz natural selection apiary and failed in the spring of 2011.

Varroa infestations on adult bees (phoretic phase) were evaluated by washing 200 to 300 bees in 100% methanol. Analyses of Varroa in sealed brood (percent brood infestation and number of Varroa per cell) and natural Varroa drop onto sticky boards was also monitored. For molecular analyses, several hundred adult worker bees were collected from the brood nest and white-eyed, pink-eyed and dark-eyed pupae were collected from sealed brood of both S88 and G4 colonies in September 2010. Pupae and adult bees, either infested or not infested with Varroa mites, were collected. The samples were frozen in liquid nitrogen and stored at $-80\,^{\circ}\mathrm{C}$.

11.3.2 Design of a honeybee-specific peptide array

To the authors' knowledge, no phosphorylation sites have been experimentally characterized in honeybee. As such, the following procedure was performed in order to identify putative honeybee phosphorylation sites. Experimentally-determined phosphorylation sites from other organisms were downloaded from the PhosphoSitePlus [Hornbeck et al., 2004, 2012] and Phospho.ELM [Diella et al., 2004, 2008, Dinkel et al., 2011] databases, and were combined into a single file. These included sites from organisms such as human, rat, mouse, cow, and *Drosophila melanogaster* (the closest honeybee relative for which phosphorylation sites are known). Phosphorylation sites were represented as 15-mer peptides, with the phosphorylated residue in the center and seven residues on either side. The honeybee proteome was constructed as follows. First, all of the honeybee proteins from UniProt (671 proteins) and GenBank (12,050 proteins) were downloaded. Second, the honeybee genome [Honeybee Genome Sequencing Consortium, 2006] was downloaded in the form of 16,501 contigs, and genes (along with their translations) were predicted using the program GeneMark.hmm [Lukashin and Borodovsky, 1998], giving 27,730 predicted proteins. Proteins from these three sources were then combined to create a final honeybee proteome consisting of 40,451 proteins. Using the DAPPLE program [Trost et al., 2013a], the 15-mer peptides from PhosphoSitePlus and Phospho.ELM were searched using BLAST against the honeybee proteome to find homologous sites. DAPPLE produced a table designed to facilitate the process of selecting honeybee peptides for inclusion on the array. Each row of the

output table corresponded to a phosphorylation site from PhosphoSitePlus or Phospho.ELM. In addition to the sequence of the best hit in the honeybee proteome, the table contained the number of sequence differences between the query peptide and the honeybee peptide, with honeybee peptides having few sequence differences being preferred. The table also included the position (e.g., Y128) of the phosphoacceptor residue for both the query peptide and the hit peptide, with honeybee peptides where the position was similar for both query and hit being preferentially selected. In addition, peptide sequences contained within proteins from UniProt or GenBank were preferred over those from proteins predicted by GeneMark.hmm. Using the above criteria, this list was manually curated to select appropriate honeybee phosphorylation sites for inclusion on the array. Peptides were selected that represent phosphorylation events associated with a broad spectrum of signaling pathways, but with specific emphasis on proteins and processes associated with innate immunity. A total of 299 peptides were ultimately selected. Each of these peptides was spotted in triplicate within each block. Further, each block was printed in triplicate, providing nine technical replicates for each peptide. Peptide synthesis, array spotting and quality control were performed as a commercial service (JPT Peptide Technologies, Berlin, Germany).

11.3.3 Kinome analysis

Application of the peptide arrays was based upon a previously reported protocol with modifications [Määttänen et al., 2013]. Briefly, individual frozen whole bees were placed in a sealed plastic bag in the presence of 300 μl of lysis buffer. The bees were struck repeatedly with a rubber mallet and the suspension was centrifuged at 10,000 $\times g$ for 10 min. Supernatants were used for kinome analysis.

11.3.4 Data analysis

The dataset for each array contained the signal intensities associated with the nine technical replicates for each of the 299 peptides for the whole body extracts of honeybee pupae or adults either uninfested or infested with Varroa mites. Those treatments were labelled "G4-" (susceptible and uninfested), "G4+" (susceptible and infested), "S88-" (resistant and uninfested), and "S88+" (resistant and infested). Kinome data were processed through PIIKA 2, a pipeline for processing kinome array data [Li et al., 2012, Trost et al., 2013b], with the following study specifics.

Consistency of technical replicates

For each peptide within a given array, a χ^2-test was performed to determine whether the degree of variability among the technical replicates for that peptide was greater than would be expected by chance. Any peptide that had a P-value according to the χ^2-test of less than 0.01 was considered to be inconsistently phosphorylated among the technical replicates.

Treatment-treatment variability analysis and pathway analysis

For each peptide, a paired t-test was used to compare its normalized signal intensity values under a treatment condition with those under a control condition. Three tests were performed for each peptide: G4+ versus G4-, S88+ versus S88-, and G4- versus S88-. Peptides with significant (P-value < 0.10) changes in phosphorylation were identified. This level of significance was chosen to retain as much data as possible in order to facilitate subsequent pathway analysis [Li et al., 2012]. Pathway and gene ontology (GO) analysis was performed as described previously [Kindrachuk et al., 2011, Määttänen et al., 2013] using InnateDB [Lynn et al., 2008].

Cluster analysis

The pre-processed data were subjected to hierarchical clustering and principal component analysis (PCA) to cluster peptide response profiles across arrays. Only peptides that were consistently phosphorylated among the technical replicates for all arrays were included in the clustering analysis. For each consistently-phosphorylated peptide on a given array, the average was taken over the nine replicates before performing clustering. For hierarchical clustering, the distance metric used was (1 − Pearson correlation), while the linkage method used was that of McQuitty (1966). Subsets of peptides that could discriminate between resistant and susceptible bees were identified as described previously [Trost et al., 2013b].

11.3.5 Virus detection

Bees were stored at −80 °C until RNA was extracted. Individual pupa were placed in small plastic bags, pulverized on dry ice, and solubilized in 700 μl Trizol (Invitrogen Canada, Burlington, ON). RNA was purified using RNeasy Mini-columns (Qiagen Canada Inc., Mississauga, ON) and RNA concentration quantified with an Agilent 2100 Bioanalyzer using RNA 6000 Nano kits (Agilent Technologies Canada Inc., Mississauga, ON). RNA pellets were re-suspended in DEPC water and converted to cDNA using qScript cDNA Supermix (Quanta Biosciences, Gaithersburg, MD). qRT-PCR was performed using PerfeCta SYBR Green Supermix for IQ (Quanta Biosciences) on a BioRad IQ5 thermocycler. Deformed wing virus was detected using primers CAGTAGCTTGGGCGATTGTT (forward) and AGCTTCTGGAACGGCAGATA (reverse) [Cox-Foster et al., 2007]. Israeli acute paralysis virus was detected using primers GCGGAGAATATAAGGCTCAG (forward) and CTTGCAAGATAAGAAAGGGGG (reverse) [Di Prisco et al., 2011]. Kashmir bee virus was detected using primers GATGAACGTCGACCTATTGA (forward) and TGTGGGTTGGCTATGAGTCA (reverse) [Cox-Foster et al., 2007]. The presence of a single PCR product of the expected size was confirmed in 2% agarose gels (Invitrogen). Detection of DWV, IAPV, and KBV was performed using an end-point PCR protocol with Phusion polymerase (New England Biolabs, Whitby, ON) with amplification at 98 °C for 30 s, then 30 cycles of: 98 °C for 10 s, 60 °C for 15 s, and 72 °C for 20 s followed by 20 s at 72 °C. Amplified products were visualized with ethidium bromide staining of 2% agarose gels. The real time cycling protocol for quantification of DWV was 95 °C for 2 min, then 40 cycles of 95 °C for 15 s, 60 °C for 30 s, and 72 °C for

30 s, followed by a melt curve to confirm amplification of a single product.

11.4 Results

11.4.1 Characterization of Varroa mite susceptible and resistant bee phenotypes

Varroa mite infestation was quantified yearly between 2007 and 2011 for the resistant (S88) colony and in 2010 for the susceptible (G4) colony (Figure 11.1A). In 2009, the average Varroa infestation rates for S88 remained below 10 per hundred bees (PHB) but ranged as high as 19 PHB. In 2010, 8 samples were analyzed between May and October showing an average infestation of 3 to 5 PHB in the S88 colony. Adult bee samples with and without Varroa were sampled in September for kinome analyses, when phoretic mite levels were 4 PHB (Figure 11.1A). S88 died in April 2011 with a Varroa mite population of 9 PHB after a colony lifespan of 58 months. This colony resisted Varroa mite population growth throughout its lifetime, although significant levels of Varroa mites persisted in the colony from establishment. High levels of phoretic Varroa were detected in May 2010 in G4 and reached as high as 67 PHB. Varroa mite population growth was very rapid in this colony (Figure 11.1A). Adult bees with and without Varroa were sampled for kinome analyses when phoretic Varroa populations were highest (September 2010). G4 died in October with a lifespan of 17 months.

These resistant and susceptible colonies were further defined by evaluating Varroa infestation in the sealed brood at the same time as adult bee samples were collected for molecular analyses. Honeybee colonies during September in Western Canada decrease brood rearing and the adult population begins to decline. Varroa increase migration into the brood, and brood Varroa levels can quickly increase. Scoring sealed G4 brood cells ($n = 500$) revealed that 88%, 84% and 70% of the white-eyed, pink-eyed and dark-eyed pupae, respectively, were Varroa-infested (Figure 11.1B). The phoretic mite levels on adult G4 bees (67 PHB) was similar to the infestation rate for dark-eyed pupae. In contrast, S88 brood infestation levels were much lower, with dark-eyed pupae infestation levels dropping to 17% from 44% and adult phoretic levels to 4 PHB (Figure 11.1B). These results imply that S88 resists Varroa population growth by removing Varroa from the brood. In addition, fewer Varroa per cell were detected in dark-eyed pupae and pre-emergent pupae in S88 than G4 at July 2010 sampling dates. G4 showed 2.7 ± 2.0 Varroa per cell ($\pm$ standard error of the mean, $n = 70$), and S88 showed 1.5 ± 1.0 Varroa per cell ($n = 9$).

11.4.2 Development of a bee-specific peptide array

The bee-specific peptide array was developed using the DAPPLE program [Trost et al., 2013a] as described in Materials and Methods. DAPPLE predicted nearly 10,000 phosphorylation events within the honeybee proteome. Of the predicted phosphorylation events, approximately 0.6% were exactly conserved over a peptide of 15 amino acids (seven residues flanking each side of the phosphoacceptor site) (Supplementary

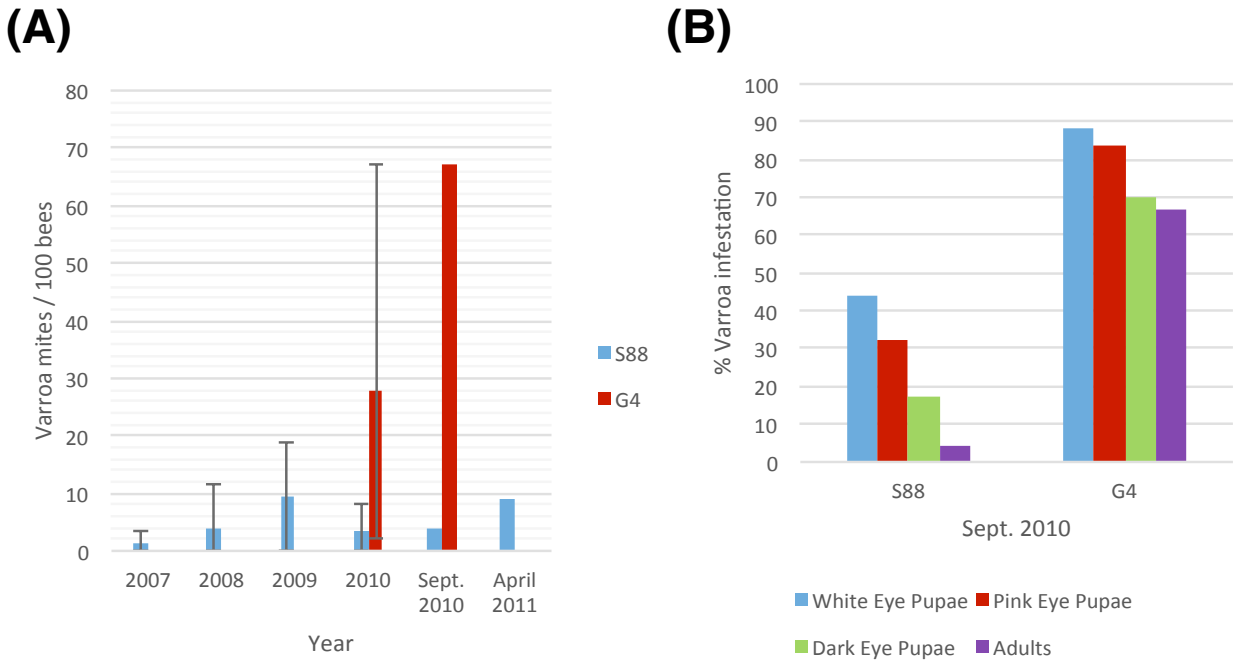


Figure 11.1: Quantification of Varroa mite infestation of G4 and S88 bees. (A) Average phoretic Varroa infestations per hundred bees in S88 and G4 colonies. Bars show the range of yearly phoretic Varroa infestations in S88 (2007 to 2010) and G4 (2010). (B) Percent Varroa infestation in sealed brood at different stages of development. Over 500 sealed brood cells were analyzed for each colony and scored for presence of Varroa.

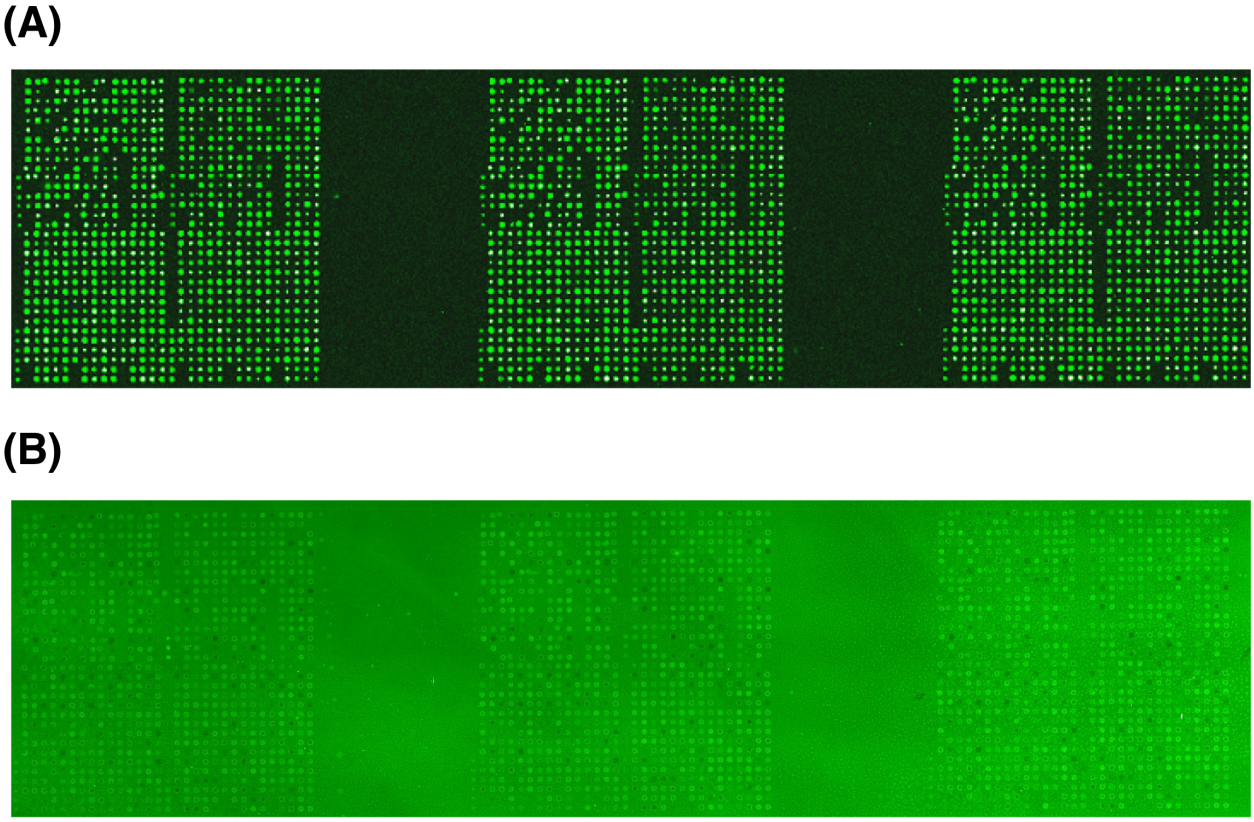


Figure 11.2: Printing and validation of the bee-specific peptide array. (A) The arrays were printed by a commercial provider (JPT Peptide Technologies, Berlin, Germany). For each array, each spot was printed in triplicate within each block. Each block was then printed in triplicate, for a total of nine technical replicates of each peptide. This image, taken as a quality control step in array production, illustrates the consistency and reproducibility of peptide spotting. (B) An image of a data scan of a representative array used for analysis of a whole-bee sample. A clear and consistent pattern of peptide phosphorylation is apparent across the three printed blocks.

Table G.1). The low degree of conservation highlights the importance of developing species-specific arrays as opposed to simply translating commercially available arrays across species.

From this panel, 299 unique phosphorylation events were selected using the criteria described in Materials and Methods. Peptides were selected to represent phosphorylation events associated with a broad spectrum of signaling pathways (to facilitate novel discovery) but with emphasis on pathways and processes associated with insect innate immunity. A GenePix Array List (GAL) file containing the exact layout and content of the array used in this study is provided (Supplementary File 1).

An image highlighting the format of the arrays as well as the consistency and reproducibility of peptide spotting is presented (Figure 11.2A). An image of a data scan of a representative array used for analysis of a whole-bee lysate is also provided (Figure 11.2B). All of the arrays used in this study were of comparable quality with respect to the clarity and consistency of peptide phosphorylation.

11.4.3 Kinome profiling of bee phenotype at different developmental stages

Uninfested bees ($n = 3$) of each phenotype (G4 and S88) were considered at each of three developmental stages (pink-eyed pupae, dark-eyed pupae and adult). In each case, kinome analysis was performed with lysate extracted from the whole organism. Morphologically, there was a clear distinction between each developmental stage. There was, however, no obvious difference in bee morphology when comparing between G4 and S88 within each development stage. The relationships among the 18 kinome datasets were evaluated through hierarchical clustering (Figure 11.3A) and three-dimensional PCA (Figure 11.3B). There was a clear indication of distinct developmentally-specific kinome profiles. Further, within each developmental stage, there was strong evidence of distinct kinome profiles for the G4 and S88 bees, indicating that Varroa mite susceptibility or resistance is reflected at the level of signal transduction.

11.4.4 Phosphomarkers of Varroa mite susceptibility in dark-eyed pupae

The ability of the arrays to detect distinct kinome profiles (kinotypes) corresponding to each phenotype suggests that the arrays may represent a valuable tool for identification of kinase activity biomarkers that are associated with resistance or the response to Varroa mite infestation. Specifically, the bee-specific peptide array, representing 299 phosphorylation events, was able to discriminate between each developmental stage, and between the two phenotypes within each developmental stage (Figure 11.3).

To determine whether smaller sets of peptides could also discriminate between the phenotypes, the peptide subset analysis described by Trost et al. [2013b] was performed on the bees at the dark-eyed pupae stage. This procedure was used to identify subsets of peptides having the property that, when samples were clustered using these peptides, bees of the same phenotype clustered together as closely as possible. This was done for peptide subsets of size 3 to 200. For subsets of selected cardinalities (5, 10, 25, 50, 100, 150, and 200), the random tree analysis described by Trost et al. [2013b] was performed to determine whether that set of peptides discriminated between the susceptible and resistant phenotypes better than would be expected by chance. It was discovered that subsets of as few as five peptides could discriminate the resistant and susceptible bee phenotypes with a high degree of confidence (P-value < 0.001) (Table 11.1). Given this, it may be possible to create a smaller, more targeted array that could provide unique kinomic profiles for each phenotype. Such a peptide subset could serve as a minimal array of practical value for screening bees within breeding programs as well as for assurance of phenotype in the sales and marketing of commercial bees.

11.4.5 Kinomic responses of susceptible and resistant dark-eyed pupae to Varroa mite challenge

Kinome profiles were determined for individual dark-eyed pupae ($n = 3$) of both the G4 and S88 colony phenotypes in the presence and absence of Varroa mite infestation. Hierarchical clustering analysis of the kinome data demonstrated distinct clustering on the basis of Varroa mite susceptibility, indicating distinct

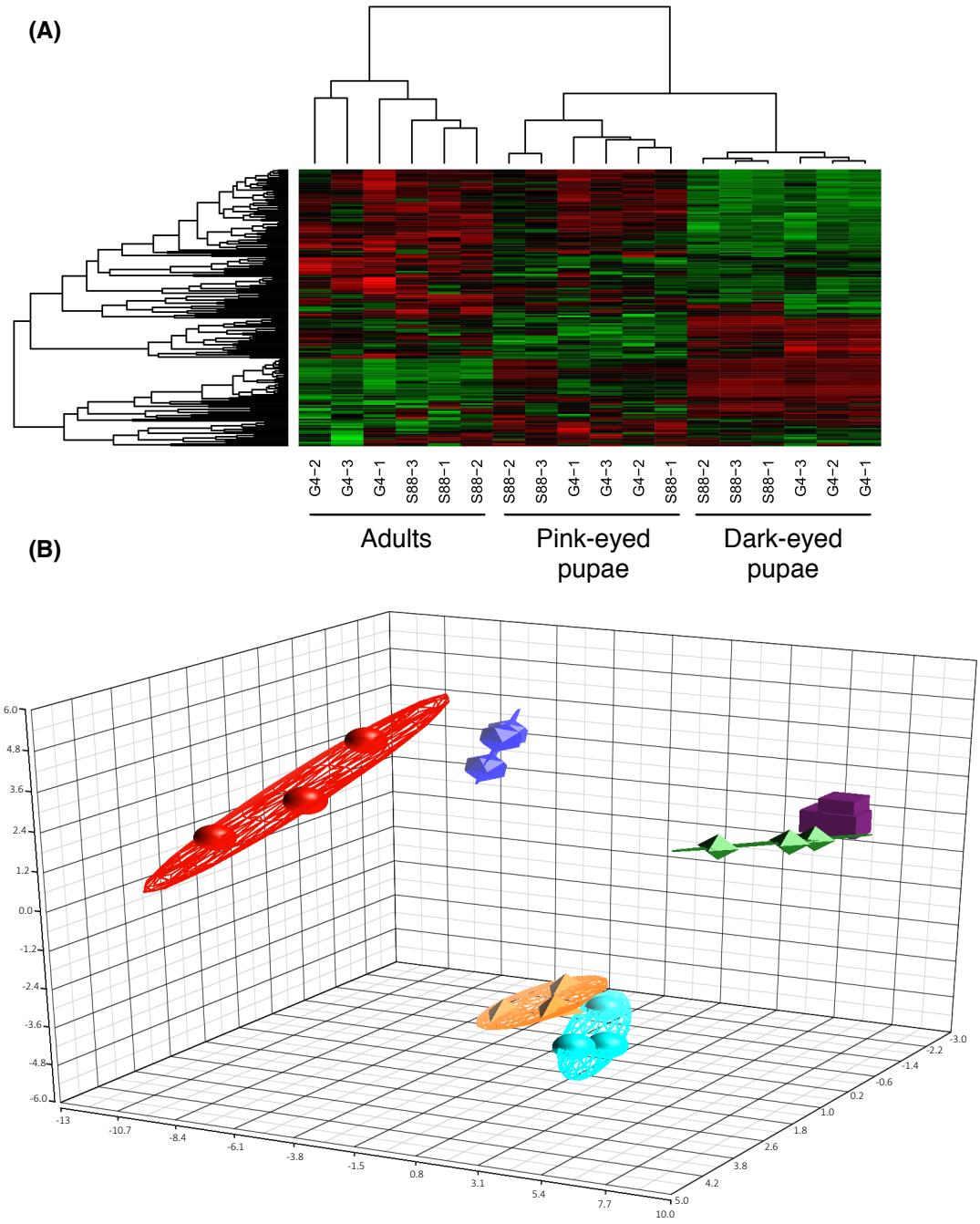


Figure 11.3: Clustering of the kinome profiles of bees of different phenotypes at different developmental stages. (A) Hierarchical clustering of kinome datasets. (1 − Pearson correlation) was used as the distance metric, while McQuitty linkage was used as the linkage method. Each column represents the kinome activity of individual bees ($n = 3$/treatment). The kinome profiles of the bees segregated first by developmental stage and then largely by colony phenotype (S88: resistant; G4: susceptible). Colors indicate the average (over 9 intra-array replicates) normalized phosphorylation intensity of each target, with red indicating greater amounts of phosphorylation and green indicating lesser amounts of phosphorylation. (B) Principal component analysis. The first three principal components are shown. The points are as follows: red, adult G4; dark blue, adult S88; green, dark-eyed G4; purple, dark-eyed S88; orange, pink-eyed G4; light blue, pink-eyed S88. The proportions of variance explained by the first, second, and third principal components were 29.1%, 15.3%, and 7.5%, respectively.

Table 11.1: Ability of subsets of peptides to discriminate susceptible and resistant bees at the dark-eyed pupae stage. Subsets of peptides were determined that best differentiated susceptible and resistant dark-eyed pupae. For selected subsets, a statistical test [Trost et al., 2013b] was used to determine whether those peptides could discriminate between the two phenotypes better than would be expected by chance. The first column of the table contains the size of the peptide subset, while the second column contains the P-value associated with this statistical test.

Number of peptides	P-value
200	0.0006
150	0.0002
100	0.0007
50	0.0001
25	0.0002
10	0.0004
5	0.0004

patterns of phosphorylation-mediated signal transduction within the two phenotypes (Figure 11.4A). This was confirmed with PCA, in which distinct clustering of samples corresponding to the phenotypes was also observed (Figure 11.4B). For both hierarchical clustering and PCA, there was further sub-clustering based on the infestation status of the samples within the susceptible phenotype. This sub-clustering was not observed within the resistant samples, except for one S88 infested pupae which showed some overlap with the susceptible G4 phenotype. These observations imply Varroa parasitism induced a more pronounced change in intracellular physiology within Varroa susceptible bees compared to resistant bees.

11.4.6 Cellular mechanisms of Varroa mite susceptibility

The kinome data were interrogated to define the biological differences between bee phenotypes at the dark-eyed pupae stage of development. Many peptides were differentially phosphorylated between phenotypes or treatments. For instance, in the uninfested samples of each phenotype, there were 153 peptides (over half of the peptides on the array) for which there were significant (P-value < 0.1) differences in phosphorylation between the phenotypes. This is consistent with resistance to Varroa mite infestation being a complex and multi-faceted process.

Specific consideration of these differentially phosphorylated peptides from the perspective of gene ontology and pathway overrepresentation analysis revealed a number of points of biological difference between uninfested bees of the resistant and susceptible phenotypes (Table 11.2 and Supplementary Table G.2), between infested and uninfested bees of the susceptible phenotype (Table 11.3 and Supplementary Table G.3), and between infested and uninfested bees of the resistant phenotype (Table 11.4 and Supplementary Table G.4). When comparing uninfested bees from the two phenotypes, there were no clear differences in pathways and

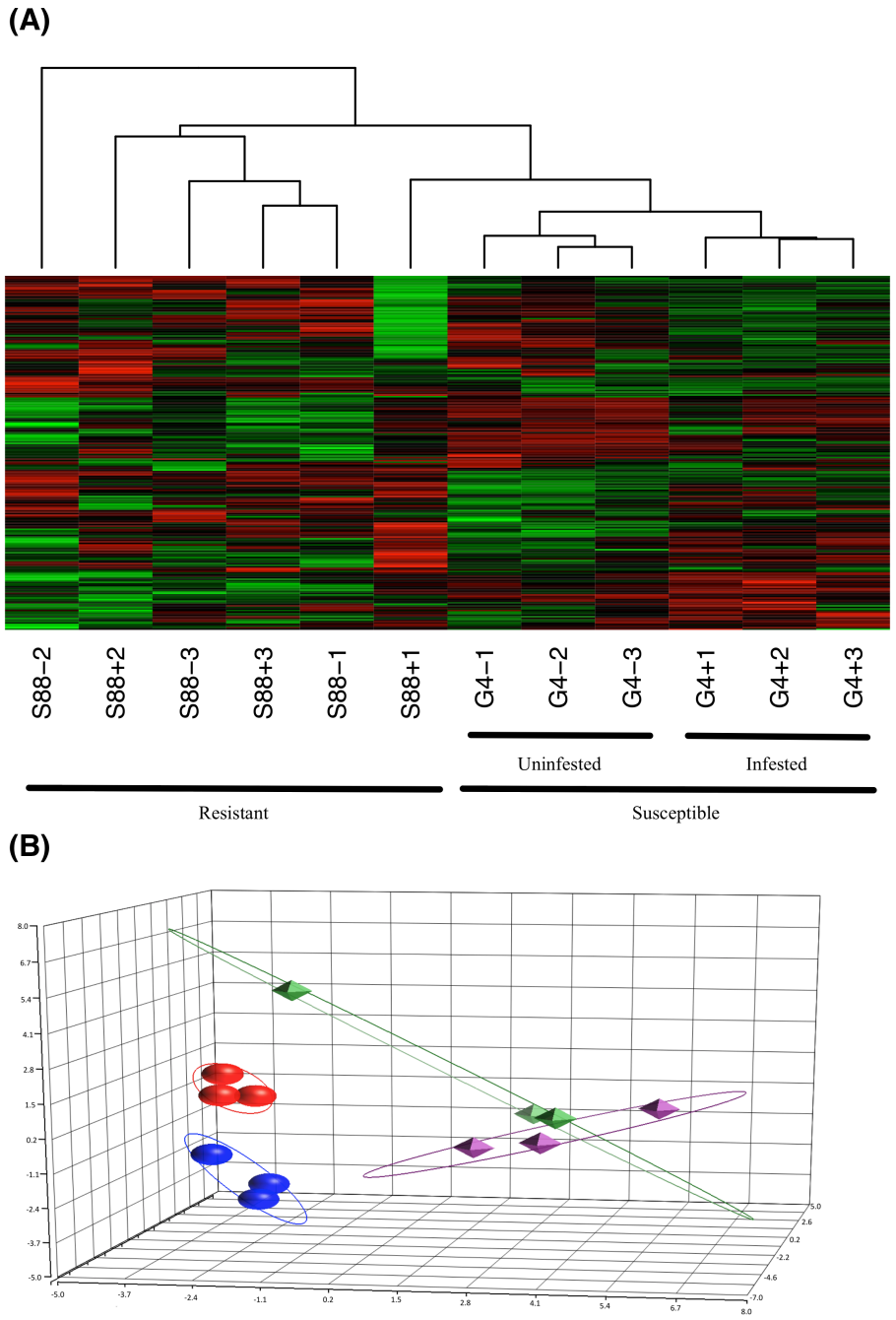


Figure 11.4: Clustering of the kinome profiles of dark-eyed pupae of different phenotypes and infestation statuses. (A) Hierarchical clustering of kinome datasets. (1 − Pearson correlation) was used as the distance metric, while McQuitty linkage was used as the linkage method. Each column represents the kinome activity of individual pupae ($n = 3$/treatment). For the most part, cluster analysis first segregated kinome profiles by colony phenotype (S88: resistant; G4: susceptible), and then segregated G4 pupae by presence or absence of Varroa infestation. (B) Principal component analysis. The first three principal components are shown. Separation of the samples on the basis of phenotype is clearly observed, with further distinction within the susceptible, but not resistant, samples on the basis of infestation status. The points are as follows: red, G4+; dark blue, G4-; green, S88+; purple, S88-. The proportions of variance explained by the first, second, and third principal components were 22.5%, 14.8%, and 11.2%, respectively.

Table 11.2: Gene ontology analysis of uninfested resistant and susceptible dark-eyed pupae (S88-/G4-). Based on levels of differential expression or phosphorylation, InnateDB [Lynn et al., 2008] can predict upregulated or downregulated pathways that are consistent with the experimental data. Pathways are assigned a P-value based on the number of proteins present for a particular pathway. The numbered columns are as follows: 1, total genes uploaded for that pathway; 2, number of genes up-phosphorylated; 3, P-value for up-phosphorylation; 4, number of genes down-phosphorylated; 5, P-value for down-phosphorylation.

Category	Name	ID	1	2	3	4	5
Biological	Cell cycle arrest	GO:0007050	5	5	0.040	0	1
process	Response to peptide hormone stimulus	GO:0043434	4	4	0.078	0	1
	ATP biosynthetic process	GO:0006754	4	0	1	4	0.03
	Positive regulation of neuron apoptosis	GO:0043525	4	0	1	4	0.03
	Cytoskeleton organization	GO:0007010	6	1	0.99	5	0.05
Cellular	Cell surface	GO:0009986	7	6	0.082	1	0.98
component	Golgi apparatus	GO:0005794	7	1	0.99	6	0.02
	Plasma membrane	GO:0005886	33	14	0.96	19	0.04

processes associated with immune function (Table 11.2 and Supplementary Table G.2). An interesting exception is that within the G4 pupae, there was a trend toward the down-regulation of innate immunity (P-value < 0.1) in response to Varroa mite infestation (Table 11.3). Down-regulation of innate immune processes in response to Varroa mite infestation was not observed in the resistant phenotype (Table 11.4).

11.4.7 Detection of secondary viral infections

For bees of both phenotypes, at the dark-eyed pupae stage of development and in the absence of Varroa mites, there was a shared presence of detectable, but low levels of DWV (Figure 11.5). However, in the presence of Varroa mites there was an approximately ten thousand fold increase in DWV RNA relative to the Varroa mite-free pupae (Figure 11.5). There was also no detectable IAPV and KBV RNA in pupae of both phenotypes, regardless of the presence or absence of mite infestation (data not shown). These observations support the hypothesis that Varroa mites serve as a vector for virus transmission and that both phenotypes experience equal levels of viral infection following mite infestation. This observation supports the conclusion that kinotypic differences between pupae from the two phenotypes reflect differences in host responses to the Varroa mite and not viral infection.

The presence of immunosuppression was suggested by kinome data analysis of susceptible bees at the dark-eyed pupae stage of development. If this immunosuppression persists throughout the life of a bee, then the ability of bees to counter further infection by secondary pathogens may be compromised. Consistent with this hypothesis, screening for two additional viral bee pathogens, IAPV and KBV, confirmed higher rates of

Table 11.3: Gene ontology analysis of susceptible dark-eyed pupae (G4+/G4-). For details, see the caption of Table 11.2.

Category	Name	ID	1	2	3	4	5
Biological	Transport	GO:0006810	4	0	1	4	0.081
process	Innate immune response	GO:0045087	27	9	0.945	18	0.098
	Cell cycle	GO:0007049	9	1	0.99	8	0.03
	DNA repair	GO:0006281	5	0	1	5	0.04
	Mitotic cell cycle	GO:0000278	5	0	1	5	0.04
	Glycolysis	GO:0006096	7	6	0.031	1	0.99
	Phosphatidylinositol-mediated signaling	GO:0048015	4	4	0.039	0	1
	Multicellular organismal development	GO:0007275	6	5	0.064	1	0.992
Cellular	Nucleoplasm	GO:0005654	22	7	0.95	15	0.106
component	Plasma membrane	GO:0005886	30	17	0.099	13	0.94
	Golgi apparatus	GO:0005794	8	1	0.99	7	0.05
	Integral to membrane	GO:0016021	8	6	0.081	2	0.98
	Basolateral plasma membrane	GO:0016323	4	4	0.038	0	1
Molecular	ATPase activity	GO:0016887	4	0	1	4	0.081
function	RNA binding	GO:0003723	4	0	1	4	0.081
	RNA pol. II transcription factor activity	GO:0003705	4	4	0.038	0	1
	GTP binding	GO:0005525	6	5	0.064	1	0.992

Table 11.4: Gene ontology analysis of resistant dark-eyed pupae (S88+/S88-). For details, see the caption of Table 11.2.

Category	Name	ID	1	2	3	4	5
Biological	RNA metabolic process	GO:0016070	5	1	0.99	4	0.067
process	mRNA metabolic process	GO:0016071	5	1	0.99	4	0.067
	Nerve growth factor receptor signaling	GO:0048011	7	7	0.028	0	1
	Positive regulation of apoptotic process	GO:0043065	5	5	0.082	0	1
	Peptidyl-serine phosphorylatio	GO:0018105	7	2	0.99	5	0.069
Molecular	Phosphor-transferase activity	GO:0016772	15	12	0.087	3	0.978
function	Protein kinase binding	GO:0019901	4	0	1	4	0.018
	RNA binding	GO:0003723	5	1	0.99	4	0.067
	DNA binding	GO:0003677	7	2	0.99	5	0.069
	Kinase activity	GO:0016301	14	6	0.97	8	0.094

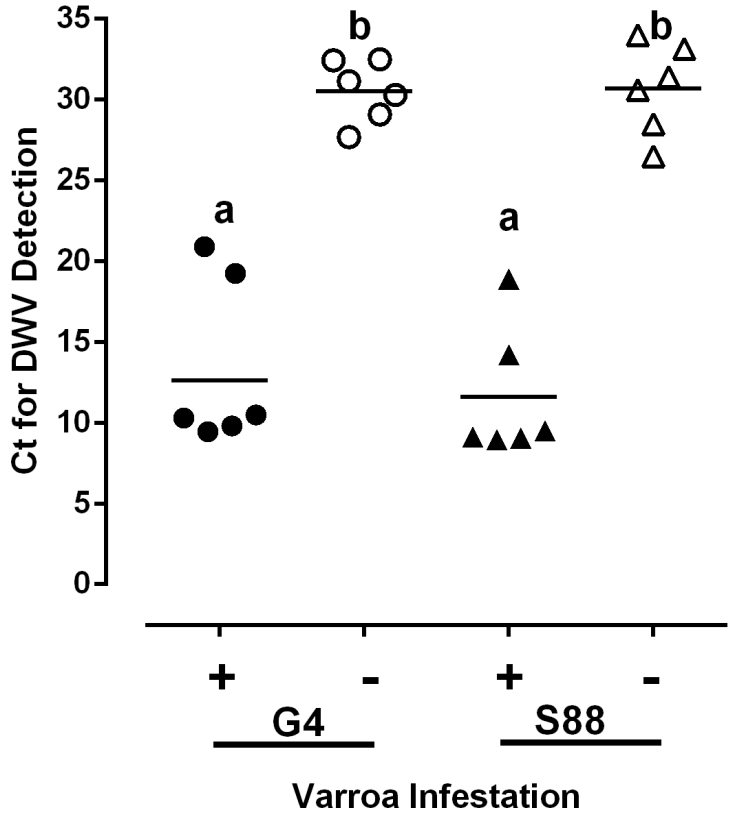


Figure 11.5: Virus presence in honeybee populations. The level of deformed wing virus (DWV) present in dark-eyed pupae was compared in the presence (+) or absence (-) of a detectable Varroa mite infestation. DWV was detected using qRT-PCR and the level of viral infection was measured as the threshold cycle (Ct) for viral RNA amplification. Ct values are inversely proportional to the abundance of viral RNA. Data presented are values for individual pupae ($n = 6$/group). Significant differences (P-value < 0.05) among treatment groups are denoted by different letters above each column.

Table 11.5: Percentage of resistant and susceptible adult bees with detectable virus. Bees (n = 20/group) were sampled in September 2010 (see Figure 11.1A). Viruses were detected using 30 cycles of amplification in qRT-PCR, and amplified products were visualized by agarose gel electrophoresis. Specific primer pairs were used to detect deformed wing virus (DWV), Israeli acute paralysis virus (IAPV), and Kashmir bee virus (KBV).

Virus	G4 (%)	S88 (%)
DWV	100	100
IAPV	60	0
KBV	15	0

infection in the susceptible adult bees in the face of Varroa mite infestation (Table 11.5).

11.5 Discussion

There is a clear and emerging priority for the ability to define global host responses at the level of phosphorylation-mediated signal transduction. As technologies advance, there is greater opportunity to apply these approaches to a broader range of species as well as samples of increasing biological complexity. Kinome analysis is often performed on cellular samples of low complexity, such as cultured cells, or purified primary cell populations, such as monocytes. Recently, there have been demonstrations of kinome analysis of samples of greater biological complexity, such as organ samples [Arsenault et al., 2013b] and intestinal tissue [Määttänen et al., 2013]. The current report, to the best of our knowledge, represents the first development of an insect-specific peptide kinome array as well as the first application of kinome analysis at the whole-organism level. The incentive to push the technology in this direction was to develop a research tool of value in the understanding of colony collapse disorder of bees. Specifically, we sought to apply the bee-specific array to populations with differing resistance to Varroa mite infestation, in the presence and absence of this critical pathogen, to provide insight into mechanisms of disease resistance as well as biomarkers for strategic bee breeding programs.

The kinome data emerging from analysis of distinct phenotypes (susceptible and resistant) at three developmental stages (pink-eyed pupae, dark-eyed pupae and adults) provided clear evidence of a phenotype-associated kinotype. As might be anticipated, each stage of development was also associated with a different global pattern of signal transduction activity. Within these development-specific patterns of clustering, there was clear evidence for distinct sub-profiles corresponding to each of the Varroa mite susceptibility phenotypes. This suggests the potential to translate the arrays into a tool that could be utilized to inform commercial aspects of bee production, such as sales and breeding. Phosphosignatures that reflect important phenotypes, such as disease resistance or production value, could be incorporated into a second generation honeybee-specific array.

In the absence of Varroa mite infestation, there were clear and consistent differences in the signaling profiles

of the susceptible and resistant bees. The magnitude of these differences suggests that resistance is a complex, multifactorial process. Interestingly, for the uninfested bees there were no obvious differences between the two phenotypes that relate to pathways or processes immediately associated with immunity. This is consistent with a previous investigation of the biological basis of Varroa mite susceptibility phenotypes through gene expression approaches, which suggested that differences in behaviour, rather than immune function, underlie Varroa resistance [Navajas et al., 2008]. The most well-defined traits associated with Varroa resistance are hygienic behavior and grooming behavior that function to maintain lower Varroa populations [Harbo and Harris, 2009, Tsuruda et al., 2012]. The S88 phenotype also showed better grooming behavior (unpublished observations). However, in our breeding efforts, it is difficult to stabilize Varroa resistant phenotypes, and the progeny of selected colony phenotypes are highly variable. Colony phenotypes can also change over time within the same colony. The survival of a resistant phenotype may be due to combinations of grooming and hygienic behavior as well as undefined mechanisms that restrict the propagation of viral pathogens. This combination of traits may be critical for bee survival in the presence of a persistent Varroa infestation. Elucidation of the mechanisms involved in this resistance to colony collapse may be critical for breeding bees able to tolerate low levels of persistent Varroa parasitism while maintaining colony health.

The responses of the two bee phenotypes to Varroa mite infestation in the current study were also investigated using pathway over-representation and gene ontology analysis. For the resistant bees, a small number of pathways were found to be activated in response to Varroa infestation. Specifically, there was robust activation of MAPK signaling, which may represent the most effective host response through induction of stress response pathways. Activation of MAPK signaling has been linked to successful management of pathogenic challenge in a number of species, including insects [Arthur and Ley, 2013]. In contrast, within the susceptible bees, there were more far-reaching consequences to Varroa mite challenge, including evidence for a down-regulation of innate immune responses.

There are conflicting opinions in the literature regarding the significance of host immunity, and the potential ability of Varroa mites to compromise host immunity. For example, some investigations have reported that Varroa mites, or virus associated with mites, compromise honeybee immunity [Gregory et al., 2005] and promote amplification of bee viruses [Yang and Cox-Foster, 2005]. From a more global perspective, a number of ectoparasites immunosuppress their vertebrate hosts and increase susceptibility to infectious disease [Yang and Cox-Foster, 2005]. Varroa mites may contribute to colony collapse by suppressing bee immunity and promoting secondary viral infections [Yang and Cox-Foster, 2005, Evans and Schwarz, 2011]. Given the conserved transmission route associated with many bee parasites, co-infection of individual bees and colonies by multiple viral pathogens is a common occurrence that can have direct and indirect interactions that may be additive, synergistic or neutral in consequences to the host [Evans and Schwarz, 2011]. Varroa mites are associated with a number of honeybee RNA viruses. In this capacity, the mites are known to contribute to colony failure both by acting as a reservoir and incubator for the viruses as well as facilitating their spread among bees [Nazzi et al., 2012]. Our work adds another layer to this synergy by suggesting that

infestation by the mite renders the bee host more susceptible to viral infection by compromising the innate immune system.

Our kinome data strongly indicate that differences in immune capabilities are likely not involved in Varroa susceptibility; rather, this phenotype may reflect primarily behavioural differences. Following Varroa mite infestation, however, the immunosuppression observed in the susceptible bees may influence their ability to counter further infestation by mites as well as secondary viral pathogens. This hypothesis is supported by greater diversity of secondary viral infections in the susceptible bees following Varroa mite infestation. This could occur at the level of the individual bees as well as the entire colony. The ultimate collapse of these colonies may represent the collective toll of these combined infections, as well as other potential stressors. This suggests that bees are not susceptible to Varroa mite infestation because they are immunocompromised; rather, they are immunocompromised because they are infested with Varroa mites. This understanding, in concert with the use of the arrays to identify appropriate biomarkers, may enable strategic breeding and management efforts to deal with the problem of Varroa parasitism and honeybee colony loss worldwide.

This initial kinome-wide analysis of honeybees has generated a number of important questions that motivate further experimental investigation. For example, more targeted investigation of the host-pathogen interaction between honeybees and Varroa mites may confirm the hypothesis that the vulnerability of the susceptible bees reflects consequences of Varroa mite infestation, as well as evidence of the molecular mechanisms involved. Unknown factors may be acting at the cellular level in Varroa resistant bees identified by natural selection (survival colonies), which may or may not be present in bees showing behavioral characteristics for expression of Varroa resistance. These factors may protect against the fatal effects associated with viruses (DWV, IAPV, KBV) vectored by Varroa, or may reduce the ability of Varroa to cause deficiencies in innate immune or stress responses. Experiments are in progress using honeybee kinome analyses to investigate these possibilities in individual bees from inbred colony lines showing varying degrees of resistance and susceptibility to Varroa. Additionally, the ability of the proposed phosphorylation-associated biomarkers of Varroa mite susceptibility should be evaluated in large-scale investigations of honeybees representing a spectrum of susceptibilities. The ability of these markers to effectively discriminate and predict this important phenotype within the context of naturally occurring variance will be important for determining the value of these markers. Ultimately, a methodology for using specific, targeted subsets of the peptide array probes (just 5 to 10 of them) to identify Varroa resistant and susceptible phenotypes needs to be developed.

11.6 Acknowledgements

This work was funded in part by grants from Saskatchewan Agriculture (Agriculture Development Fund) and the Agriculture Council of Saskatchewan to AJR and by Meadow Ridge Enterprises Ltd. PG holds a Tier I Canada Research Chair funded by the Canadian Institutes of Health Research. BT and AK received funding from the Natural Sciences and Engineering Research Council of Canada (NSERC). We thank the

Saskatchewan Beekeepers Association for administration of funds, Sanjie Jiang and Syed Shah for help with sample collection, John Pedersen and Neil Morrison for help with breeding work, Dr. Bob Danka (USDA) for critically reviewing the manuscript, and all Saskatchewan beekeepers who supported the Saskatraz breeding program.

Chapter 12

Discussion and conclusion

Since each of the main chapters of this thesis (Chapters 3-11) includes its own discussion section, this chapter mainly (though not exclusively) discusses topics that encompass more than one chapter. Specifically, Section 12.1 gives some additional detail regarding the SAPHIRE website, which hosts the three tools described in this thesis. Section 12.2 discusses how some of the work done for this thesis is applicable to areas of research beyond kinome microarrays. A comparison between PHOSFER and DAPPLE—both of which are tools designed to predict phosphorylation sites in organisms with few known sites—is given in Section 12.3. Section 12.4 is specific to DAPPLE, and discusses the relationship between the number of sequence differences between the query peptide and its best match in the target proteome, and the probability that the central residue in the hit peptide is actually phosphorylated *in vivo*. Section 12.5 reiterates the importance of good experiment design using examples encountered in the course of performing this thesis work. Section 12.6 discusses the possible application of the work done in this thesis to more biological problems. Finally, Section 12.7 discusses the collaborative nature of this thesis, and Section 12.8 contains some concluding remarks.

12.1 The SAPHIRE website

As described in Chapter 1, the SAskatchewan PHosphorylation Internet REsource (SAPHIRE) webpage contains web implementations of PHOSFER, DAPPLE, and PIIKA 2. It can be accessed via `http://saphire.usask.ca`. Common gateway interface (CGI) scripts implemented in Perl are used to process the web forms and pass parameters to the appropriate scripts. All three tools require that the user enter a valid e-mail address, and when the user's job has finished running, an e-mail is automatically sent containing a link where their results can be downloaded. Figure 12.1 contains a screenshot of the main SAPHIRE page.

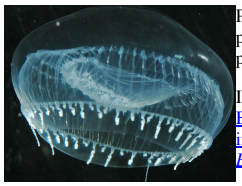
The web interface for PHOSFER (Figure 12.2) is very simple: besides the user's e-mail address, the only parameter is the protein sequences for which the user wants to make predictions. The sequences can be specified either by pasting them into a text box, or by uploading a file. PHOSFER takes a fairly short time to run; in an informal timing test, PHOSFER took approximately 3 minutes to make predictions for 1,000 protein sequences and 18 minutes for 10,000 sequences (this apparent sublinearity is due to the fact that the model files take some time to load into memory—an operation that is independent of the number of sequences used as input). Since the computational load of running PHOSFER on the server is small, the

SAPHIRE

SAskatchewan PHosphorylation Internet REsource

Welcome to the SAskatchewan PHosphorylation Internet REsource (SAPHIRE). Hosted by the University of Saskatchewan, this site currently contains three tools designed for the *in silico* analysis of phosphorylation sites.

Tool #1: PHOSFER



PHOSFER uses a novel machine-learning approach in order to predict phosphorylation sites in soybean proteins, and will be expanded to predict for other plants in the future.

If you use PHOSFER, please cite:
B. Trost and A. Kusalik. Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics* 29(6):686-694, 2013.

Tool #2: DAPPLE

DAPPLE is a homology-based method for predicting phosphorylation sites in an organism of interest. It uses BLAST searches of experimentally-determined phosphorylation sites in one organism (or several organisms) to predict phosphorylation sites in an organism of interest. It outputs a table containing information helpful for choosing phosphorylation sites that are of interest to you, such as the number of sequence differences between the query site and the hit site, the location of the query site and the hit site in their respective intact proteins, and whether the corresponding intact proteins are reciprocal BLAST hits (and thus predicted orthologues).

If you use DAPPLE, please cite:
B. Trost, R. Arsenault, P. Griebel, S. Napper, and A. Kusalik. DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites. *Bioinformatics* 29(13):1693-1695, 2013.

Tool #3: PIIKA 2

PIIKA 2 is a tool for analyzing data originating from kinome microarrays.

The original version of PIIKA is described in the following paper:
Y. Li, R. J. Arsenault, B. Trost, J. Slind, P. J. Griebel, S. Napper, and A. Kusalik. A systematic approach for analysis of peptide array kinome data. *Science Signaling* 5(220):pl2, 2012.

PIIKA 2 includes many new features and also has a web-based interface. It is described in the following paper:
B. Trost, J. Kindrachuk, P. Määttänen, S. Napper, and A. Kusalik. PIIKA 2: An Expanded, Web-Based Platform for Analysis of Kinome Microarray Data. PLOS ONE 8(11):e80837, 2013.

Image credits: Sierra Blakely (PHOSFER) and Flickr users Soggydan (DAPPLE) and wildxplorer (PIIKA 2).

Figure 12.1: The SAPHIRE website.

user is not given the option to download a stand-alone version of the tool.

DAPPLE's web interface (Figure 12.3) is also quite simple, although it has three parameters (not counting the user's e-mail address) instead of one. The first parameter is the target organism—that is, the organism for which the user wants to identify phosphorylation sites. Users can either select from a predefined list of 329 organisms (the organisms whose genomes had been sequenced at the time DAPPLE was developed), or upload their own file of protein sequences that define the proteome of some organism. The second parameter is the database of known phosphorylation sites. Here, users can either use predefined data from one of the major phosphorylation site databases (PhosphoSitePlus, Phospho.ELM, P^3DB, or PhosphoGRID), or they can upload their own file. Permission was obtained from the developers of these databases to use their data for DAPPLE. The final parameter is the maximum number of results to return per query peptide. DAPPLE can take a long time to run (perhaps more than a day, depending on the size of the phosphorylation site database used and the size of the target proteome); thus, in addition to the web interface, users also have the option of downloading the software and running it on their own computers. The DAPPLE webpage contains a link to download the software.

Compared to the web interfaces for PHOSFER and DAPPLE, the PIIKA 2 interface is more complex (Figure 8.9), with five required parameters and several optional parameters. In addition, the main input file, which is a table containing the raw phosphorylation intensities for each spot on each array, must be in a specific format. Therefore, the PIIKA 2 webpage contains a link to a second page that explains how to format the main input file and choose the correct parameters. A portion of this page is shown in Figure 12.4. Like DAPPLE, PIIKA 2 gives users the option of downloading the software.

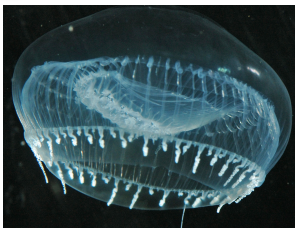
12.2 Applicability of research done for this thesis

While the work done for this mainly is largely described in the context of kinome microarrays, much of it is more broadly applicable. In particular, PHOSFER (Chapter 4) and DAPPLE (Chapter 5), which are computational methods for predicting phosphorylation sites in organisms with few known sites, would be relevant to many researchers studying phosphorylation-mediated signaling. The methodologies described in these chapters would also be relevant to the prediction of other types of PTMs (see also Section 13.1.3). In contrast to PHOSFER and DAPPLE, PIIKA (Chapter 7) and PIIKA 2 (Chapter 8) are both largely specific to kinome microarrays. However, several of the features described would also be applicable to other types of microarrays, including the statistical tests for technical and biological consistency, the visualization methods, and the techniques for the analysis of hierarchical clustering data.

12.3 Comparing PHOSFER and DAPPLE

Although PHOSFER and DAPPLE are both tools for predicting phosphorylation sites in organisms having few experimentally verified sites, each uses a fundamentally different approach. DAPPLE uses a homology-

PHOSFER



PHOSFER uses a novel machine-learning approach in order to predict phosphorylation sites in soybean proteins. To use PHOSFER, simply provide the sequence(s) in FASTA format for which you want predictions, as well as your e-mail address.

The use of PHOSFER is free for academic, non-commercial purposes. If you would like to use it for commercial purposes, please contact us.

Step #1: Input file

Please paste the sequences (multi-FASTA format) for which you want to make predictions in the box below.

Alternatively, choose a file from your local computer to upload.

Choose File no file selected

Step #2: E-mail address

Please enter your e-mail address here. Once your job is finished running, you will receive an e-mail with a link where you can download the results.

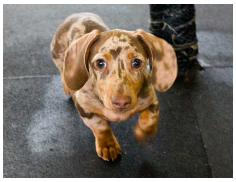
Step #3: Submit!

Submit

Image credit: Sierra Blakely

Figure 12.2: The PHOSFER web interface.

DAPPLE



DAPPLE represents an alternative method (to machine-learning approaches) to predicting phosphorylation sites in an organism of interest. It is a pipeline involving BLAST searches that uses experimentally-determined phosphorylation sites in one organism (or several organisms) to predict phosphorylation sites in an organism of interest. It outputs a table in tab-deliminated text format (which can also be easily imported into a spreadsheet program like Excel), which contains various information helpful for choosing phosphorylation sites that are of interest to you, such as the number of sequence differences between the query site and the hit site, the location of the query site and the hit site in their respective intact proteins, whether the corresponding intact proteins are reciprocal BLAST hits (and thus predicted orthologues), and so on.

The following is a web interface to DAPPLE. If you would instead like to run DAPPLE on your own machine, you may download it here. This .zip file includes instructions for setting up DAPPLE.

Step #1: Target organism

Please select the target organism (for which you want to predict phosphorylation sites) from the list below.

Acromyrmex echinatior (Panamanian leafcutter ant) [taxid=103372]

Alternatively, choose a proteome file (multi-FASTA format) from your local computer to upload.

Choose File no file selected

Step #2: Known phosphorylation site data

Please select the dataset of known phosphorylation sites that you would like to use. You may choose from four different databases:

- PhosphoSitePlus (Contains phosphorylation sites from a variety of organisms, mostly mammals)
- phospho.ELM (Contains phosphorylation sites from a variety of organisms)
- P^{3}DB (Contains phosphorylation sites from plants)
- PhosphoGRID (Contains phosphorylation sites from *Saccharomyces cerevisisae* only)

We would like to thank the developers of these databases for kindly giving their permission to use their data with DAPPLE.

PhosphoSitePlus

Alternatively, choose a file from your local computer to upload. Click here to view a sample of the required format.

Choose File no file selected

Note: As some of these databases are quite large, DAPPLE can take a significant amount of time to run (more than a day for PhosphoSitePlus). Please run only one instance of DAPPLE at a time. If you wish, you may also use the standalone version of DAPPLE, for which you can run as many simultaneous instances as you like (or your machines will allow).

Step #3: Max results per query

Please select the maximum number of results to return per known phosphorylation site.

1

Step #4: E-mail address

Please enter your e-mail address here. Once your job is finished running, you will receive an e-mail with a link where you can download the results.

Step #5: Submit!

Submit

Image credit: Flickr user Soggydan.

Figure 12.3: The DAPPLE web interface.

PIIKA 2 input guide

This page describes the various input files and parameters associated with using PIIKA 2.

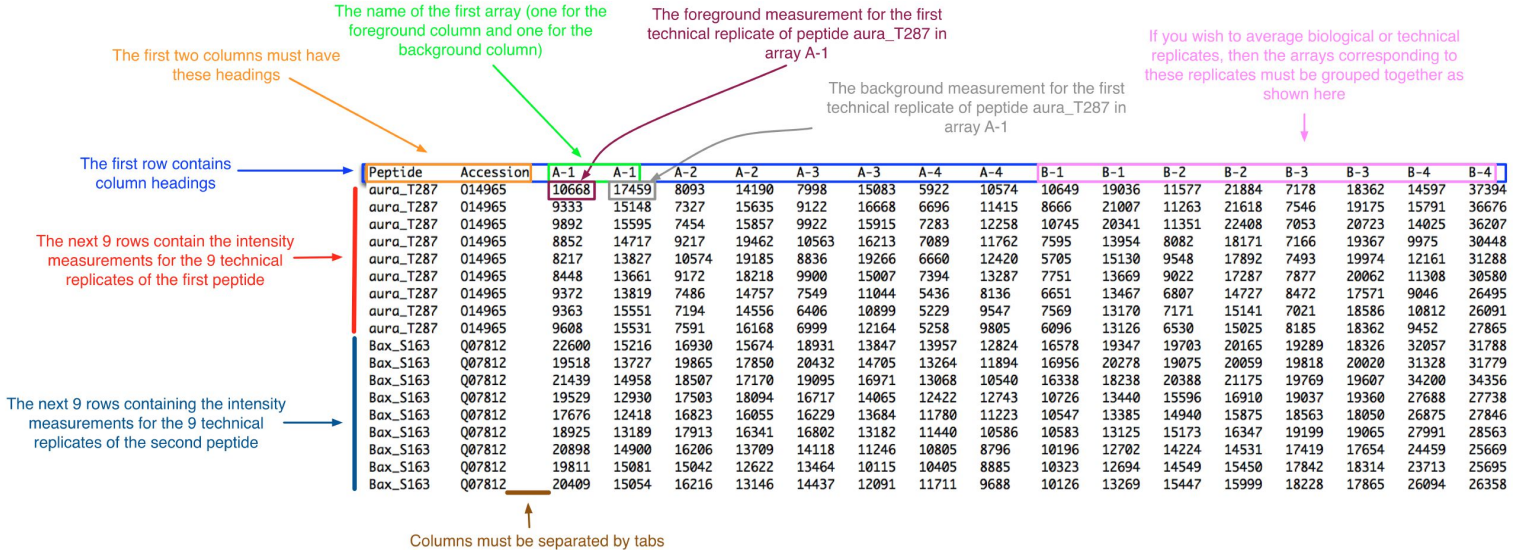
Step #1: Input files

Main input file (required)--- The main input file must be a file in tab-delimited text format. The format of the input file must follow these rules.

- The first row contain column headings.
  - The first two column headings must be "Peptide" and "Accession", respectively.
  - Subsequent column headings correspond to arrays in your experiment, and should be labeled in groups of two, with the same name for each, to correspond with the foreground and background intensity readings for a single array.
    - *For example, in the sample input file, the first array is called "A-1", so the third and fourth column headings are each labeled "A-1".*
- If your arrays contain *n* technical replicates for each peptide, then each subsequent set of *n* lines (after the header line) must contain the data for those replicates, one replicate per line.
  - *For example, the arrays used to produce the data in the sample input file contained 9 technical replicates, so lines 2-10 of the file contain the data for the first peptide, lines 11-19 contain the data for the second peptide, and so on.*
- Within a line (other than the first line):
  - The first two columns contain the peptide name and accession number of the protein corresponding to that peptide, respectively (as suggested by the column headings).
  - The next two columns contain the foreground and background intensity values, respectively, for the first array; the next two columns contain these values for the second array, and so on.

Important note: If all of the arrays in your experiment correspond to different treatments/controls, then the order of the columns (except for the first two) is unimportant. However, if you have arrays from, say, multiple animals that all received the same treatment (biological replicates), then the data for the arrays corresponding to the same treatment must be in adjacent columns. Although the 4 arrays corresponding to each subject in our sample data were not grouped together for analysis purposes, suppose that we did want to group them together (for, say, clustering). Then all of the arrays corresponding to subject A must appear together, and the same for the other subjects, as shown in the sample file.

The following figure illustrates the use of the above rules using a portion of the sample input file.



Peptide	Accession	A-1	A-1	A-2	A-2	A-3	A-3	A-4	A-4	B-1	B-1	B-2	B-2	B-3	B-3	B-4	B-4
aura_T287	014965	10668	17459	8093	14190	7998	15083	5922	10574	10649	19036	11577	21884	7178	18362	14597	37394
aura_T287	014965	9333	15148	7327	15635	9122	16668	6696	11415	8666	21007	11263	21618	7546	19175	15791	36676
aura_T287	014965	9892	15595	7454	15857	9922	15915	7283	12258	10745	20341	11351	22408	7053	20723	14025	36207
aura_T287	014965	8852	14717	9217	19462	10563	16213	7089	11762	7595	13954	8082	18171	7166	19367	9975	30448
aura_T287	014965	8217	13827	10574	19185	8836	19266	6660	12420	5705	15130	9548	17892	7493	19974	12161	31288
aura_T287	014965	8448	13661	9172	18218	9900	15007	7394	13287	7751	13669	9022	17287	7877	20062	11308	30580
aura_T287	014965	9372	13819	7486	14757	7549	11044	5436	8136	6651	13467	6807	14727	8472	17571	9046	26495
aura_T287	014965	9363	15551	7194	14556	6406	10899	5229	9547	7569	13170	7171	15141	7021	18586	10812	26091
aura_T287	014965	9608	15531	7591	16168	6999	12164	5258	9805	6096	13126	6530	15025	8185	18362	9452	27865
Bax_S163	Q07812	22600	15216	16930	15674	18931	13847	13957	12824	16578	19347	19703	20165	19289	18326	32057	31788
Bax_S163	Q07812	19518	13727	19865	17850	20432	14705	13264	11894	16956	20278	19075	20059	19818	20020	31328	31779
Bax_S163	Q07812	21439	14958	18507	17170	19095	16971	13068	10540	16338	18238	20388	21175	19769	19607	34200	34356
Bax_S163	Q07812	19529	12930	17503	18094	16717	14065	12422	12743	10726	13440	15596	16910	19037	19360	27688	27738
Bax_S163	Q07812	17676	12418	16823	16055	16229	13684	11780	11223	10547	13385	14940	15875	18563	18050	26875	27846
Bax_S163	Q07812	18925	13189	17913	16341	16802	13182	11440	10586	10583	13125	15173	16347	19199	19065	27991	28563
Bax_S163	Q07812	20898	14900	16206	13709	14118	11246	10805	8796	10196	12702	14224	14531	17419	17654	24459	25669
Bax_S163	Q07812	19811	15081	15042	12622	13464	10115	10405	8885	10323	12694	14549	15450	17842	18314	23713	25695
Bax_S163	Q07812	20409	15054	16216	13146	14437	12091	11711	9688	10126	13269	15447	15999	18228	17865	26094	26358

What is the best way to make a file like this? That depends on the format of the files produced by your image analysis software. If you are familiar with programming and/or UNIX utilities, you can write a program that takes those files as input and outputs a file in the format specified above. If you do not have this expertise, the best method would probably be to construct the file in Excel (or the spreadsheet program of your choice). To use Excel, follow these general guidelines:

1. Open a new Excel file.
2. Save the file in **tab-delimited text (.txt)** format (call the file main_input_file.txt).
3. Enter "Peptide" in cell A1 and "Accession" in cell B1.
4. Import the first file from your image analysis software (corresponding to the first array in your experiment) into Excel as a separate spreadsheet.
5. Copy and paste the names of the peptides and the accession numbers from this file into columns A and B, respectively, of main_input_file.txt starting at row 2.
6. Copy and paste the foreground and background measurements in this file into columns C and D, respectively, of main_input_file.txt starting at row 2
7. Put appropriate column headings in cells C1 and D1.
8. Repeat steps 4, 6, and 7 for each array, putting the new data into successive columns, making sure that the order of the peptides are identical for each array.

Figure 12.4: Part of the PIIKA 2 input guide.

based method wherein the set of proteins from the organism of interest is used as a BLAST database and known phosphorylation sites from other organisms are used as queries. In contrast, PHOSFER uses a machine-learning technique based on random forests (although homology does play an indirect role in the derivation of instance weights for the training data).

Given that DAPPLE and PHOSFER have identical goals but use different approaches, the question arises: how do they compare in terms of predictive performance? Chapter 4 allows an initial answer to this question to be derived, as it gives the sensitivity and specificity of PHOSFER and allows the sensitivity of DAPPLE to be estimated.

The sensitivity and specificity of PHOSFER are given explicitly in Table 4.3. At a specificity of 0.99, PHOSFER had sensitivities of 0.337, 0.167, and 0.204 for S, T, and Y sites, respectively. At specificity 0.95, the corresponding sensitivity values were 0.545, 0.324, and 0.245.

The sensitivity of DAPPLE can be estimated from Table 4.2, which shows the level of phosphorylation site conservation between soybean and the other organisms used in the PHOSFER study. The H^1_{Bk} column contains the percentage of known phosphorylation sites in soybean that had homologous sites in the proteome of each organism. This information gives a rough idea of the sensitivity that would be realized if DAPPLE were used with soybean as the target proteome and known phosphorylation sites from one of the other organisms as the database. For example, if 20% of the phosphorylation sites in soybean were also found in organism X, then using known phosphorylation sites from X to identify sites in soybean should result in 20% of the soybean sites being identified, giving a sensitivity of 0.2. Table 4.2 shows that the projected sensitivity of DAPPLE varies greatly depending on the relatedness of the organism from which the known phosphorylation sites are derived. For instance, the percentage of sites in soybean that had an equivalent site in *Arabidopsis* (the closest relative to soybean among the organisms tested) was 36% for S sites, 37.5% for T sites, and 58.3% for Y sites. This suggests that DAPPLE would have a sensitivity of 0.36, 0.375, and 0.583 for S, T, and Y sites, respectively, if soybean were used as the target proteome and known phosphorylation sites from *Arabidopsis* were used as queries. Table 4.2 predicts that rice, another plant used in the PHOSFER study, would have lower sensitivity, while distantly-related organisms like human, mouse, and yeast would have approximate sensitivity values of less than 0.05 for all three site types.

While Table 4.2 provides insight into the sensitivity of DAPPLE, determining its specificity is more difficult. Gathering negatives (amino acid residues that are not phosphorylated) presents a challenge for predictors based on machine learning, since it is difficult to say with confidence that a particular site does not undergo phosphorylation *in vivo*. However, this problem is even more acute for DAPPLE. This is because predictions made by DAPPLE are evidence-based rather than pattern-based. In contrast to machine-learning methods, which predict phosphorylation sites based on the pattern of residues surrounding the potential site itself, DAPPLE uses as evidence the fact that the analogous residue in a homologous protein in another organism is known to be a phosphorylation site. This is different than in machine-learning methods, where the evidence involved (patterns characterizing the amino acid composition of known phosphorylation sites)

is more indirect. Thus, it is difficult to identify DAPPLE-predicted sites for which there is any confidence that the site is, in actuality, not a phosphorylation site.

How, then, does DAPPLE compare to PHOSFER in terms of predictive performance? Given that the specificity of DAPPLE is difficult to determine, assume for the moment that phosphorylation events are well-conserved among related organisms. In this case, a specificity of 0.95 may plausibly be used to compare the sensitivity of DAPPLE with that of PHOSFER. At this specificity level, PHOSFER had substantially better sensitivity for S sites (0.545 versus 0.36), while DAPPLE had much better sensitivity for Y sites (0.583 versus 0.245). DAPPLE and PHOSFER exhibited similar sensitivities for T sites (0.375 and 0.324, respectively). Given their respective prediction strategies, the performance of DAPPLE should be positively related to the level of phosphorylation site conservation, while the performance of PHOSFER should be negatively related to the degeneracy of the amino acids surrounding a particular phosphorylated residue. Therefore, the aforementioned comparisons suggest that S sites may have more predictable patterns (boosting the performance of PHOSFER) among the three residue types, while Y sites may be better conserved (improving the performance of DAPPLE). It should be emphasized that these observations are somewhat speculative, as they are based on the assumption of DAPPLE having approximately 0.95 specificity. In addition, while these observations may apply to plants, they may not generalize to all organisms. It is possible, for instance, that Y sites are more well-conserved in plants than S and T sites, but that this is not true in mammals.

12.4 The relationship between number of sequence differences and the probability that a peptide contains a phosphorylation site

The output of DAPPLE is a table where each row represents a peptide X (usually 15 amino acids in length) whose central residue is a known phosphorylation site, and each column contains information about that peptide's best match Y in the target proteome. A user that is designing a kinome microarray (or predicting phosphorylation sites for some other purpose) would want to select peptides Y that have a high probability p of containing a central residue that is indeed phosphorylated *in vivo*.

One of the most important columns for choosing such peptides is "Sequence differences", which contains the number of sequence differences between X and Y. Let n denote the value of this column. When selecting phosphorylation sites from the DAPPLE output table, a critical assumption is that n and p are inversely related; in other words, the fewer sequence differences, the greater the likelihood that the central residue in Y is a real phosphorylation site. Under this assumption, the user should choose as many peptides as possible for which n is small. This assumption seems reasonable given the strong relationship between conservation of sequence and conservation of function among proteins in general. However, users of DAPPLE should be aware that it is, nonetheless, an assumption: to the author's knowledge, there does not exist a study that has explicitly examined the relationship between n and p. Of particular interest would be a threshold t such that a certain (large) percentage of peptides Y with $n \leq t$ contain a central residue that is phosphorylated

in vivo. Determining the value of p for larger values of n (say, between 5 and 7) would also be valuable, as it would shed light on the validity of choosing weaker matches from DAPPLE output.

12.5 The importance of good experiment design

Good experiment design is a critical component of any study involving kinome microarrays. The goal of most such studies is to determine the effects of one or more treatments (for example, infection with a virus or administration of a drug) on cellular signaling pathways. This often involves comparing the responses of animals that have been exposed to the treatment to those that have not. However, as described in Section 2.2.4, natural biological variation among the animals can overwhelm signals relating to the treatments being investigated.

During the course of the research done for this thesis, the importance of good experiment design has been highlighted multiple times. As one example, consider the study described in Chapter 10, whose goal was to determine the cellular pathways affected by MAP infection in calves. The intestine of each calf was surgically isolated and divided into three compartments; two of these were infected with MAP, while the third was left uninfected. After a period of time, tissue samples were extracted from each compartment and analyzed using kinome microarrays. When the kinome profiles of the different tissue compartments from the calves were subjected to hierarchical clustering (Figure 10.3A), no obvious pattern was observed. However, when the normalized intensity values for each uninfected compartment were subtracted (separately) from the infected compartments from the same animal, the clustering pattern was consistent with the type of immune response that the animals exhibited (Figure 10.3B). Specifically, the two calves exhibiting cell-mediated immune responses (which is favourable for eliminating MAP infection) clustered together, as did the two calves exhibiting antibody immune responses (which is not favourable). This implies that some relationship may exist between kinome responses and the type of immune response generated. This observation was made possible because the experimental design allowed biological subtraction to be performed (see Section 2.2.4), which eliminates the effect of biological variation. Had the experiment been designed such that the entire intestine of each calf was either infected or uninfected, this pattern would likely not have been evident.

The need for good experiment design was also highlighted by work done as part of a research contract with a commercial company (due to confidentiality provisions in the contract, the precise nature of the research will not be described). The company provided samples extracted from several infected organisms, as well as several uninfected organisms. Ideally, samples would have been taken from the same individual both pre- and post-infection; unfortunately, each sample was, in fact, taken from a different individual. When the data were subsequently analyzed, few meaningful patterns could be discerned. This can likely be attributed to the natural biological variation among the individuals tested having a greater impact on the kinome profiles than the presence or absence of infection. Given previous experience, more meaningful results would likely have been obtained if both infected and uninfected samples were taken from the same individual.

12.6 Applying kinome microarrays to biological problems in different species

In this thesis, three papers were presented that described the application of kinome microarrays to biological problems (Chapters 9, 10, and 11). While all three of these studies used PIIKA 2 to analyze the data from the kinome arrays, only one (Chapter 11) used computationally predicted phosphorylation sites in order to design an array (the peptides on the human/pig array used for Chapter 9 were drawn directly from phosphorylation site databases, while the bovine sites for the arrays used in Chapter 10 were identified using the manual method described by Jalal et al. [2009]). However, we have also used computational prediction of phosphorylation sites to design arrays in other studies. In a project not otherwise described in this thesis, DAPPLE was used to create a chicken-specific array. This array was used to examine how hot or cold temperature stress (often experienced by chickens being transported for slaughter) affects meat quality, as well as the chickens' kinome profiles. DAPPLE has also been used to design hamster and horse arrays; projects utilizing these arrays are ongoing. It is hoped that other researchers will use DAPPLE to aid the study of other species as well. As PHOSFER is currently limited to soybean (see also Section 13.1.2), it has not yet been put to practical use; however, once it has been extended to other organisms, it should become more valuable for the design of species-specific kinome microarrays.

12.7 Collaborations for this thesis

Much of the work presented in this thesis involved a significant amount of collaboration, and I am extremely grateful for the knowledge, expertise, and hard work of every one of my co-authors. Given its heavy reliance on the availability of biological data, this thesis would not have been possible without them.

Each main chapter (3–11) contained a description of the contributions of each co-author; however, given the collaborative nature of this thesis, it seems valuable to clarify my contributions to the thesis as a whole. Except for feedback and advice from Dr. Kusalik, Chapters 3 and 4 are entirely my work. I performed all of the programming and design for the software described in Chapter 5, and also wrote the manuscript, while co-authors helped develop the methodology, contributed ideas and feedback, and helped revise the manuscript. Except for the final selection of peptides for the honeybee array, Chapter 6 is entirely my work. Yue Li, with support from Dr. Kusalik, Dr. Napper, Dr. Arsenault, and Dr. Griebel, developed the initial methodology and implementation of PIIKA (Chapter 7). I helped revise the methodology, wrote parts of the paper, revised the paper, and wrote the implementation of PIIKA that was ultimately described. I contributed the majority of the ideas and methodology for PIIKA 2 (Chapter 8), developed the web implementation, and wrote the majority of the paper. The wet-lab work for Chapter 9 was done by Dr. Napper's lab. I developed most of the data analysis methods used, while additional data analysis was performed by Dr. Kusalik, Dr. Kindrachuk, and Dr. Napper. Several authors, including me, collaborated in writing and revising the paper.

The data presented in Chapter 10 were generated by Dr. Määttänen, with significant work in experiment design by Dr. Griebel and Dr. Napper. Dr. Määttänen performed the majority of the data analysis; however, I analyzed the kinome array data, helped with other aspects of data analysis, and wrote the sections of the paper relating to the analysis of the kinome data. Finally, Chapter 11 was a collaboration among several authors: Dr. Robertson and colleagues performed the bee breeding and characterized the bee phenotypes, and Dr. Napper's lab performed the wet-lab experiments. I performed much of the work in designing the honeybee-specific arrays, helped with data analysis, wrote parts of the paper, and contributed substantial revisions to the paper.

12.8 Conclusion

The phosphorylation of proteins plays an integral role in cellular signaling processes, which in turn determine the physiology of the cell. Given their complexity, achieving a broad understanding of cellular signaling networks requires the ability to analyze many phosphorylation reactions simultaneously. While still a relatively new technology, kinome microarrays have been used for this purpose in a number of studies. However, they are truly useful only if appropriate peptides can be selected for inclusion on an array, and if the data resulting from the arrays can be analyzed in a valid and meaningful way. This thesis described the development of tools that facilitate these two tasks. Specifically, DAPPLE and PHOSFER were developed for the former task, while PIIKA and PIIKA 2 were developed for the latter task. Of course, the development of these tools is significant only if they are used to analyze real biological problems. Thus, this thesis included three chapters that concentrated on biological applications of kinome arrays. Manuscripts are currently in preparation that describe additional studies involving the use of the aforementioned tools to study biological problems. It is hoped that the tools developed as part of this thesis will prove useful for addressing many biological problems in the future, and also serve as a basis for further work on computational aspects of kinome microarrays.

Chapter 13

Future work

Beyond this thesis, there are many avenues for further work on computational aspects of kinome microarray experiments. These can be divided into two categories: those that involve the design of kinome arrays (Section 13.1), and those that relate to the analysis of data resulting from the arrays (Section 13.2).

13.1 Design of kinome microarrays

This section discusses possibilities for future work that relate to the design of kinome microarrays. Specifically, Section 13.1.1 describes a potential method for increasing the computational efficiency of DAPPLE, while Section 13.1.2 discusses the modification of PHOSFER to predict for organisms other than soybean. Finally, Section 13.1.3 discusses a topic indirectly related to the design of kinome microarrays—the adaptation of DAPPLE and PHOSFER to predict for other types of post-translational modifications.

13.1.1 Using faster database search algorithms for DAPPLE

As described in Chapter 5, DAPPLE requires the execution of many BLAST searches. Using the same notation as in Chapter 5, let X represent a peptide containing a known site from a phosphorylation site database. DAPPLE first performs a BLAST search using X as the query and the target proteome as the database. Call its best hit Y; further, let X' and Y' denote the full proteins corresponding to X and Y, respectively. In order to determine whether X' and Y' are orthologues, DAPPLE uses the reciprocal BLAST hits method. This necessitates two additional BLAST searches: X' must be searched against the target proteome, and Y' must be searched against the proteome of the organism encoding X'. Thus, if the phosphorylation site database contains n sites, then $3n$ BLAST searches may be required. In practice, however, the number of searches will be less than $3n$. If X has no hit in the target proteome, then neither BLAST search for ascertaining orthology needs to be done. If X does have a hit, but the best hit when X' is searched against the target proteome is not Y', then the second BLAST search (Y' searched against the organism encoding X') is unnecessary, since it has already been determined that X' and Y' are not reciprocal BLAST hits. Nonetheless, since the value of n can be large (the PhosphoSitePlus database currently contains 255,759 known sites), DAPPLE can take a significant amount of time to run on a single processor (around a day for a single target proteome, depending on the value of n, the number of proteins in the target proteome,

and other variables).

Given that sequence database searches comprise the majority of DAPPLE's running time, it would be desirable to speed up this portion of the procedure. One option is to run BLAST on multiple cores/processors. Another is to use a faster search algorithm than BLAST. One alternative is called PAUDA [Huson and Xie, 2014], whose authors claim is thousands of times faster than BLAST. Unfortunately, it is also substantially less sensitive; in its authors' tests, only 33% of the query sequences that had a database hit using BLAST also had a hit using PAUDA. This level of sensitivity makes PAUDA inappropriate for use in DAPPLE. A better alternative may be RAPSearch 2 [Ye et al., 2011, Zhao et al., 2012], which works by using a reduced amino acid alphabet—instead of using all 20 amino acids when performing the similarity search, amino acids with similar chemical properties are considered to be the same. For instance, Lys and Arg are grouped together because each has a positive charge. The use of a reduced alphabet enables longer seeds to be chosen (see also Section 2.3.1, which explains how seeds are used in BLAST), which results in a substantial reduction in the number of seeds that must be extended, as well as a corresponding reduction in compute time. The authors of RAPSearch 2 report that it is 20-90× faster than BLAST while having only slightly reduced sensitivity.

13.1.2 Extending PHOSFER to predict for organisms other than soybean

Currently, PHOSFER only implements a phosphorylation site prediction model for one organism—soybean. This organism was chosen as a test case because the number of known phosphorylation sites for soybean was small enough to fit the purpose of PHOSFER (predicting for organisms having few known sites), but large enough that the model for PHOSFER could be meaningfully compared to a model created using only known sites from soybean.

In Chapter 4, it was suggested that future work could involve implementing models for additional organisms. Since a proof of concept has already been given for PHOSFER, additional organisms need not satisfy the "few known sites, but not too few" criterion mentioned above. In fact, the fewer the known sites for a given organism, the more benefit (in terms of increased prediction accuracy) can likely be obtained from using known phosphorylation sites from other organisms as training data. Despite the increasing use of mass spectrometry for the high-throughput identification of phosphorylation sites, there remain many plants of economic importance for which few experimentally-determined phosphorylation sites exist. For example, as of November 2013, P^3DB contained just 818 sites from rapeseed (of which canola is one cultivar), 115 sites from corn, and 33 sites from potato (Table 1.1). Although the title of Chapter 4 specifically mentions plants, this was done mainly to differentiate PHOSFER from other predictors, most of which were trained using only mammalian sites; there is nothing plant-specific about the algorithm used in PHOSFER. Therefore, it could potentially be applied to non-plant organisms having few known sites, such as chicken (364 sites in PhosphoSitePlus) and sheep (12 sites).

Another interesting question that could be addressed is, "Does using phosphorylation sites from other organisms provide a benefit when many sites are known in the organism of interest?" On one hand, using

data from other organisms would increase the amount of information used, which is usually desirable in machine-learning problems. On the other hand, these extra data may increase the amount of noise. As it is not clear whether the benefit of more information would outweigh any added noise, empirical tests would need to be done using an organism with many known sites (such as human or mouse) in order to answer this question.

13.1.3 Extending PHOSFER and DAPPLE to predict for other post-translational modifications

While PHOSFER and DAPPLE are designed to predict for one particular type of PTM—phosphorylation—both could be adapted to predict for other PTMs as well. While this would not be useful for designing kinome arrays, it could be applied to the design of arrays for other PTMs. DAPPLE would require almost no modification; the only thing that would need to change is the database of sites (i.e., instead of a database containing phosphorylation sites, one would use a database corresponding to the PTM of interest). On the surface, PHOSFER would also be relatively easy to modify, with the only required change being the training data. However, it is possible that some aspects of the algorithm itself may not be appropriate for other PTMs. For instance, the machine-learning model used in PHOSFER uses a peptide length of 15. Although this has been shown to be an appropriate length for phosphorylation sites [Biswas et al., 2010], it could be too long or too short for other PTMs. An empirical evaluation of predictive performance would be needed to determine appropriate parameter values.

Of course, making a predictor for any PTM requires the existence of a database containing known instances of that PTM. Thankfully, databases exist for several different PTMs. For instance, O-GLYCBASE is a database of glycosylation sites [Hansen et al., 1998, Gupta et al., 1999], while both UbiProt [Chernorudskiy et al., 2007] and hUbiquitome [Du et al., 2011] are online databases dedicated to ubiquitination. The Compendium of Protein Lysine Modifications (CPLM) contains experimentally-determined sites for many types of modifications to the amino acid lysine [Liu et al., 2011]. In addition to phosphorylation sites, PhosphoSitePlus contains databases of several other PTMs, including acetylation, methylation, and sumoylation [Hornbeck et al., 2012].

Relative to phosphorylation, fewer sites are known for other PTMs. This may be because fewer exist in nature (phosphorylation being one of the most widespread PTMs), and because they are less well-studied. For instance, the latest version of O-GLYCBASE contains data for only 242 glycoproteins, while PhosphoSitePlus contains just 789 sumoylation sites. Since both DAPPLE and PHOSFER are designed for situations in which there is an *organism-specific* deficiency in known sites (that is, both tools assume that a greater number of sites are known in organisms other than the one of interest), they may not perform as well for post-translational modifications in which the number of known examples is small among all organisms. This is particularly true of DAPPLE: if n sites exist for a given PTM among all organisms, then no more than n sites could be predicted in the organism of interest. Thus, the potential usefulness of DAPPLE is limited when n is

small. Since PHOSFER makes predictions based on a model built by analyzing patterns in the training data, it could potentially predict a far greater number of sites than DAPPLE for small n. The weakness of PHOSFER in such situations would likely be its accuracy, as it is difficult to build an accurate model when the size of the training set is small (in contrast, the accuracy of DAPPLE is independent of n).

13.2 Analysis of kinome microarray data

This section proposes ideas for future work on the computational analysis of data from kinome microarrays. Section 13.2.1 suggests comparing the efficacy of additional transformation and normalization methods in PIIKA 2. Section 13.2.2 explains the potential value of creating standards for the reporting of kinome array data, as well as an online database for storing them, similar to what already exists for DNA microarrays. While kinome array data are currently used to identify already-known signaling pathways that are differentially modulated between a treatment condition and a control condition, it would be even more interesting (and challenging) to use such data to identify novel signaling pathways; this is proposed in Section 13.2.3. Section 13.2.4 suggests that systematic error could potentially be reduced by improving the layout of technical replicates on the arrays. While hierarchical clustering and PCA are currently used in PIIKA 2, other clustering methods could also be investigated; these are discussed in Section 13.2.5. The addition of multiple hypothesis testing correction to PIIKA 2 is discussed in Section 13.2.6. Ideas for improving the peptide subset analysis in PIIKA 2 are given in Section 13.2.7, while the possibility of investigating different databases for pathway analysis is detailed in Section 13.2.8. Finally, the generation of artificial kinome microarray data is discussed in Section 13.2.9.

13.2.1 Comparing different transformation and normalization methods

One of the steps in the PIIKA 2 pipeline is normalization, which converts the raw phosphorylation intensity data to a more usable form by ensuring that all values are positive and bringing the data from multiple arrays (each of which may be subject to systematic biases) onto the same scale. Currently, the normalization method used in PIIKA 2 is VSN [Huber et al., 2002], which was found to be the best method among several tested (Chapter 7). However, there are many normalization methods that could be investigated beyond those tested in Chapter 7, such as the variance stabilizing transformation [Durbin et al., 2002]. This project would involve identifying several additional normalization methods, implementing those methods in PIIKA 2, and then testing them. As in Chapter 7, the efficacy of each normalization method could be measured using two criteria: the statistical properties of the data after normalization, and the effect of the method on the ability of PIIKA 2 to correctly identify differentially modulated signaling pathways. The normalization method identified to be the best could then be used as the default method in PIIKA 2.

13.2.2 Databases and standards for kinome microarray data

Online databases of biological information have been recognized as important for the research community for many years. The journal *Nucleic Acids Research* (NAR), for instance, has published over 20 annual issues devoted to biological databases [Fernández-Suárez and Galperin, 2013]. The rate of database development has increased substantially over the years—the first database-specific issue published by NAR described 18 databases, while the 2013 issue covered 88 new databases and 88 updates to existing databases [Fernández-Suárez and Galperin, 2013]. The Molecular Biology Database Collection, a compendium of online databases described in NAR, now includes over 1500 entries [Fernández-Suárez and Galperin, 2013]. It is important to note that this figure includes only databases described in NAR—many have been described in other journals as well. The importance of online bioinformatics databases is such that established publishers are creating journals devoted specifically to this subject; one example is the journal *Database*, which is published by Oxford University Press. Indeed, portions of this thesis would have been difficult or impossible without databases of biological sequences, such as UniProt [Apweiler et al., 2004, Boutet et al., 2007, UniProt Consortium, 2013], as well as databases of phosphorylation sites, such as PhosphoSitePlus [Hornbeck et al., 2004, 2012].

Aside from databases of protein and nucleic acid sequences, some of the most-used databases are those that store information from DNA microarray experiments. There are two major repositories for this information: the Gene Expression Omnibus (GEO) [Edgar et al., 2002, Barrett et al., 2005, 2007, 2009, 2011, 2013] and ArrayExpress [Brazma et al., 2003, Parkinson et al., 2005, 2007, 2009, 2011, Rustici et al., 2013], each of which currently contains data from around a million arrays [Barrett et al., 2013, Rustici et al., 2013]. Both databases feature tools for searching, displaying, visualizing, and even analyzing the data contained therein. It has been formally recommended that users of DNA microarrays submit their data to one of these databases [Ball et al., 2004].

Closely linked to DNA microarray databases is the minimum information about a microarray experiment (MIAME) [Brazma et al., 2001] standard, which dictates what information should be reported about a DNA microarray experiment. Many journals require studies involving DNA arrays to be MIAME-compliant [Functional Genomics Data Society, 2010], and the ArrayExpress database automatically verifies the compliance of submitted data [Rustici et al., 2013].

No data reporting standards or online databases currently exist for kinome microarray data. However, the success of the MIAME standard, as well as of online databases for storing DNA array data, suggest that similar facilitates may prove beneficial for kinome microarrays. A standard analogous to MIAME could easily be developed—many of the requirements, such as information regarding the nature of the samples, would remain the same, while other aspects, such as the requirement to list the nucleotide sequences on the array, have obvious analogues for kinome arrays. The development of a MIAME-like standard for kinome arrays would increase the ability to understand, evaluate, and reproduce experiments. Similarly, having an online database dedicated to kinome microarray data could have several benefits. Such a database would facilitate the comparison of data from disparate experiments; for example, meta-studies could be performed

in the same vein as the one performed by Lukk et al. [2010] for DNA microarrays. It would also allow more people—even those not involved in the generation of kinome microarray data—to examine published data, find meaningful patterns, and draw biological conclusions.

13.2.3 Identifying novel signaling pathways

A great deal of information is known regarding cellular signaling pathways, particularly in well-studied organisms like human and mouse. Given the incredible complexity of these pathways, however, many are undoubtedly yet to be discovered, particularly in less well-studied organisms.

Given that kinome microarrays provide information regarding the upregulation or downregulation of known signaling pathways in response to different treatments, it would be interesting to investigate whether they could help identify novel signaling pathways. A machine-learning approach could be used, with the positive training data being comprised of sets of peptides on the array that are known to be part of the same signaling network, and the negative training data being comprised of randomly-generated peptide sets. The goal would then be to identify patterns characteristic of the sets of peptides corresponding to signaling pathways, but not found in the randomly-generated sets. While simple in concept, this would likely be a difficult problem in practice, as the upregulation or downregulation of a given pathway is not necessarily independent of the regulation of other pathways, and a given protein can be involved in more than one pathway.

13.2.4 Characterizing the effects of spot position on intensity measurements

As described in Section 2.2, the spots on a kinome microarray are arranged in a grid. The relative positions of the spots containing the technical replicates for a given peptide sequence are important, as certain arrangements could induce systematic error. As an extreme example, it would not be desirable for all of the technical replicates corresponding to a particular peptide to be adjacent to one another.

Figure 2.7 shows the pattern in which technical replicates are found on the arrays used for the experiments described in this thesis. Specifically, the red spots represent the nine technical replicates corresponding to a particular peptide sequence. The layout of these technical replicates has some desirable properties. Specifically:

- three of the technical replicates are found in each of the three level A blocks;
- each of the technical replicates is found in a different level C block; and
- within a given level A block, the three technical replicates are found in different positions in their respective level C blocks.

However, the arrangement also has some undesirable properties:

- the three technical replicates within a given level A block have the same configuration as they do in the other two level A blocks; and
- within a given level A block, all three replicates are found in the same level B block.

As future work, an array configuration could be developed that retains the above desirable properties while eliminating the undesirable properties. More generally, the goal would be to identify a configuration that has the least potential for systematic bias. Note that simply putting each spot in a random position on the array may not be ideal, because the technical replicates for some peptides might end up in, say, the same level C block simply by chance.

13.2.5 Comparing different clustering methods

PIIKA 2 provides two clustering methods: hierarchical clustering and PCA. Given that many clustering methods exist, it may be valuable to evaluate others to determine their suitability and usefulness for clustering kinome microarray data. In this context, clustering may be applied to samples or to peptides. Since the number of objects is very different for each (hundreds of peptides versus perhaps 3-50 samples), potential clustering methods should be tested on both, as a method with good performance on a dataset with only a few objects may perform poorly on a dataset with many objects, or vice versa. The remainder of this section describes three clustering methods currently not available in PIIKA 2. One of these—t-distributed stochastic neighbor embedding—has already been evaluated (albeit informally), while the other two—self-organizing maps and fuzzy clustering—remain as future work. As all three methods are implemented in R (via the functions `tsne`, `SOM`, and `fanny`, respectively), they could easily be integrated into PIIKA 2.

T-distributed stochastic neighbor embedding

Like PCA, t-distributed stochastic neighbor embedding (t-SNE) [van der Maaten and Hinton, 2008] is a dimensionality reduction technique. Building on a previous technique called stochastic neighbor embedding (SNE) [Hinton and Roweis, 2002], t-SNE works by first converting Euclidean distances between data points to conditional probabilities. These conditional probabilities, which represent similarities between pairs of objects, are subsequently mapped into a lower-dimensional space.

The authors of t-SNE claim that it offers a number of advantages over other dimensionality reduction techniques, including greater robustness for data with certain characteristics and an improved ability to visualize both local and global structure in the data. Unfortunately, informal testing suggested that t-SNE does not perform well when clustering samples from kinome microarray experiments. Two-dimensional scatterplots of t-SNE-transformed data revealed no clearly-defined clusters; rather, all samples were spaced almost equidistant from one other. This same pattern was evident for data from several different kinome microarray experiments. Additionally, the results from t-SNE were inconsistent with those given by hierarchical clustering and PCA. Determining the ability of t-SNE to cluster peptides remains as future work.

Self-organizing maps

A self-organizing map (SOM) is a variant of an artificial neural network initially used for the semantic analysis of natural language [Kohonen, 1990]. Like hierarchical clustering, SOMs are constructed using an iterative algorithm; however, it has been argued that they lack many of the deficiencies inherent in hierarchical clustering while still being computational practical [Tamayo et al., 1999].

In building a SOM, the objects are first mapped into n-dimensional space, where n is a small number suitable for visualization (e.g., 2 or 3). A small number of nodes, which represent clusters, are initially placed at random in this space. Like hierarchical clustering, a distance metric is used; the chosen metric determines the distance between an object and a node, as well as the distance between two nodes.

Before the first iteration, the objects are placed in a randomly ordered list, with the indices starting at zero. For a given iteration i, the element with index $i \bmod M$ is chosen, where M is the number of objects. Let $f_i(N)$ represent the location of node N at iteration i, and let $N_P = \text{argmin}_N d(f_{i-1}(N), P)$—that is, N_P is the closest node to object P. Each node, including N_P, is then moved closer to P. The amount by which a given node is moved depends on two factors: the value of i (nodes move farther in earlier iterations) and the distance from the node to N_P (the smaller the value of $d(N, N_P)$, the greater the distance moved).

SOMs have been successfully used for clustering genes according to their expression patterns in DNA microarray data. For instance, Tamayo et al. [1999] used them to analyze gene expression changes in yeast cells as they progress through the cell cycle. Using 30 nodes and 50,000 iterations, the authors found that genes involved in cell cycle regulation tended to be either in the same node or in neighbouring nodes. While SOMs have been demonstrated to be useful for clustering DNA microarray data, they have not yet been used for kinome microarrays; thus, their usefulness in that context still needs to be ascertained.

Fuzzy clustering

Most clustering algorithms are described as "hard", which means that a given object can be a member of only a single cluster. In contrast, "fuzzy" or "soft" clustering methods potentially assign a given object to multiple clusters. The output of a fuzzy clustering method is a matrix M, where a given element M_{oc} denotes the degree to which object o is a member of cluster c [Kaufman and Rousseeuw, 1990]. The greater the value of M_{oc}, the greater the extent to which o belongs to c. The matrix values have the property that $\sum_c M_{oc} = 1$. Fuzzy clustering may be especially useful for clustering peptides, as there is a natural biological interpretation for a given peptide being a member of more than one cluster (i.e., the protein containing that peptide may be involved in multiple signaling pathways).

13.2.6 Multiple hypothesis testing in PIIKA 2

When comparing the degree of peptide phosphorylation between two samples using PIIKA 2, a t-test is performed for each peptide, and the resulting P-value reported. In the current version, no adjustment for

multiple hypothesis testing is made. Future work could thus involve choosing an appropriate method for handling the multiple hypothesis testing problem when analyzing kinome microarray data.

A number of such techniques have been proposed. The Bonferroni correction, which is perhaps the simplest technique, involves dividing α by the number of tests performed. However, given that kinome microarray experiments typically involve a large number of peptides (as well as a relatively small number of replicates per peptide), the Bonferroni correction may be too conservative—it is, in fact, possible that no peptides would attain a P-value less than the Bonferroni-corrected value of α for a given pair of samples.

Another technique is to adjust the desired significance level in order to maintain an acceptable estimated false discovery rate. The false discovery rate is the proportion of peptides predicted to be differentially phosphorylated that were predicted incorrectly; in terms of the quantities described in Section 2.3.2, it is equal to $\mathrm{FP}/(\mathrm{FP}+\mathrm{TP})$. Suppose that a control condition was being compared to a treatment condition on a 300-peptide kinome microarray, and that 50 substrates had P-values less than $\alpha = 0.05$. Also, let A_N represent the actual number of peptides that are not differentially phosphorylated. The estimated number of false positives could theoretically be calculated as $\alpha \times A_N$. However, A_N is typically unknown. However, if the (unrealistic) assumption is made that there are no false negatives, then the number of false positives can be estimated. Given this assumption, $\mathrm{TN} = 300 - 50 = 250$. Also, the ratio between FP and TN is $\alpha/(1-\alpha) = 0.05/0.95$. Thus, the number of false positives can be estimated as $\mathrm{TN} \times (\mathrm{FP}/\mathrm{TN}) = 250 \times 0.05/0.95 = 13.2$. The false discovery rate can then be estimated: $13.2/50 = 0.26$. If this rate is viewed as too high, then the value of α can be lowered in order to reduce it. For example, if α was reduced to 0.01, 25 substrates might attain P-values less then this threshold. The expected number of false positives would then be $(300 - 25) \times 0.01/0.99 = 2.8$, giving an estimated false discovery rate of $2.8/25 = 0.11$.

A variant of the above procedure was described by Stekel [2003]. There, the quantity $\mathrm{FP}/(\mathrm{FP}+\mathrm{TP})$, which is conventionally referred to as the "false discovery rate", is instead called the "false positive rate". In addition, in contrast to the procedure described above, Stekel estimates the number of false positives by multiplying α by the number of peptides on the array (300). This may provide for reasonably accurate estimates when the actual number of differentially phosphorylated peptides is low, but not when it is high.

Besides the above, there are many other methods for handling the multiple hypothesis testing problem. An analysis and comparison of several of these methods is given by Farcomeni [2007], and Dudoit et al. [2003] give a review of multiple hypothesis testing with specific reference to DNA microarray experiments.

13.2.7 Improving the peptide subset analysis in PIIKA 2

As described in Chapter 8, PIIKA 2 includes a feature that identifies subsets of peptides whose clustering is as consistent as possible with a particular *a priori* clustering of the samples. This feature could be enhanced in at least two ways: by replacing the current algorithm with one that can identify peptide subsets whose clustering more closely matches the hypothesized clustering, and by allowing more than one subset of a particular cardinality to be returned. Both potential improvements are described in more detail below.

In Chapter 8, a metric $\delta'(T)$ was proposed that defines how well the binary tree T corresponding to a hierarchical clustering conforms to a user's hypothesized clustering. An algorithm was presented that attempts to find peptide subsets of size n from the set full set P that induce a clustering with as large a value of $\delta'(T)$ as possible. As it would be computationally intractable to try all $\binom{|P|}{n}$ possible combinations of peptides to determine the one that maximizes $\delta'(T)$, PIIKA 2 instead uses a greedy heuristic that, while tractable, is not guaranteed to identify the optimal subset. Therefore, future work could involve finding a computationally tractable optimal algorithm, or—more likely—a different non-optimal algorithm that improves upon the current one in terms of average- or worst-case performance.

In the initial portion of the current algorithm used by PIIKA 2, the values of $\delta'(T)$ are computed for all possible sets of two peptides. The set with the greatest value of $\delta'(T)$ is used as the "seed" for the remainder of the algorithm; if multiple subsets have the same value of $\delta'(T)$, then one of these is arbitrarily selected as the seed. It is currently not known how the seed selection affects the composition of the larger subsets computed thereafter. Thus, future work could involve characterizing the effect of the seed selection procedure on the composition of the larger subsets. If it is found that the seed selection makes a substantial difference, then further work could involve modifying PIIKA 2 to report subsets resulting from different seeds.

13.2.8 Comparing different pathway databases

In this thesis, two different databases were used for pathway analysis: InnateDB [Lynn et al., 2008, Breuer et al., 2013] and the commercial product Ingenuity Pathway Analysis (IPA) [Ingenuity Systems, 2013]. InnateDB focuses primarily on pathways involved in immunity, while IPA contains a wide variety of biological pathways. In addition to these, several other pathway databases exist, including the Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY database [Kanehisa et al., 2010], MetaCyc [Caspi et al., 2008], and Reactome [Joshi-Tope et al., 2005, Croft et al., 2011].

Given the number of available pathway databases and their complexity, it would be difficult and time-consuming for a user to attain a thorough understanding of what data are present in the above databases, how they overlap, and what analysis tools each offers. Therefore, a systematic comparison of these databases would be a valuable resource. While one study compared MetaCyc with KEGG [Altman et al., 2013], to the author's knowledge there has not yet been a study comparing all of the pathway databases mentioned above. Such a comparison could involve evaluating the databases using several criteria, including:

- availability of tools for pathway visualization;
- ability to perform pathway over-representation analysis;
- user-friendliness;
- number of pathways and pathway components represented; and
- representation of different organisms.

While the analysis of data from kinome microarrays would be one potential "use case" of these databases, the use cases that would be relevant to a given user would depend on what type of data is being analyzed. Therefore, a thorough study would involve identifying a wide variety of use cases and then identifying the most appropriate database for each one.

13.2.9 Generating artificial kinome microarray data

When performing a kinome microarray experiment, it is usually the case that the "correct" answers (in terms of which peptides should be differentially phosphorylated, which pathways should be upregulated or downregulated, which samples should cluster with which other samples, and so on) are unknown. This makes it more difficult to evaluate and compare potential components of a pipeline (such as PIIKA 2) for analyzing kinome microarray data. For instance, in Chapter 7, different transformation and normalization methods were compared. While one criterion for this comparison was based on the statistical properties of the transformed data, another criterion related to how well differentially modulated signaling pathways could be identified using the transformed data. The latter task was aided by the fact that CpG, LPS, and IFN are known to activate the TLR, IL-2, and JAK-STAT pathways, respectively. However, this comparison could be improved if the full effect of CpG, LPS, and IFN on phosphorylation-mediated cellular signaling was known. As this is unlikely to be the case in the foreseeable future, an alternative would be to devise a program that generates artificial kinome microarray data. These data would need to faithfully reflect the statistical characteristics of real data, and have the property that the "correct" result of analyzing the data is known. In addition to normalization and transformation methods, the selection of other elements of the analysis pipeline, such as the pathway database (see also Section 13.2.8), could benefit from artificial data.

References

T Altman, M Travers, A Kothari, R Caspi, and P D Karp. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, 14:112, 2013.

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

R Apweiler, A Bairoch, C H Wu, W C Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, D A Natale, C O'Donovan, N Redaschi, and L S Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, 32(Database issue):D115–9, 2004.

R Arsenault, P Griebel, and S Napper. Peptide arrays for kinome analysis: new opportunities and remaining challenges. *Proteomics*, 11(24):4595–609, 2011.

R J Arsenault, S Jalal, L A Babiuk, A Potter, P J Griebel, and S Napper. Kinome analysis of Toll-like receptor signaling in bovine monocytes. *J Recept Signal Transduct Res*, 29(6):299–311, 2009.

R J Arsenault, Y Li, K Bell, K Doig, A Potter, P J Griebel, A Kusalik, and S Napper. *Mycobacterium avium* subsp. *paratuberculosis* inhibits gamma interferon-induced signaling in bovine monocytes: insights into the cellular mechanisms of Johne's disease. *Infect Immun*, 80(9):3039–48, 2012.

R J Arsenault, Y Li, P Maattanen, E Scruten, K Doig, A Potter, P Griebel, A Kusalik, and S Napper. Altered Toll-like receptor 9 signaling in *Mycobacterium avium* subsp. *paratuberculosis*-infected bovine monocytes reveals potential therapeutic targets. *Infect Immun*, 81(1):226–37, 2013a.

R J Arsenault, S Napper, and M H Kogut. *Salmonella enterica* Typhimurium infection causes metabolic changes in chicken muscle involving AMPK, fatty acid and insulin/mTOR signaling. *Vet Res*, 44(1):35, 2013b.

J S C Arthur and S C Ley. Mitogen-activated protein kinases in innate immunity. *Nat Rev Immunol*, 13(9): 679–92, 2013.

C A Ball, A Brazma, H Causton, S Chervitz, R Edgar, P Hingamp, J C Matese, H Parkinson, J Quackenbush, M Ringwald, S-A Sansone, G Sherlock, P Spellman, C Stoeckert, Y Tateno, R Taylor, J White, and N Winegarden. Submission of microarray data to public repositories. *PLoS Biol*, 2(9):E317, 2004.

J P Bannantine, J F J Huntley, E Miltner, J R Stabel, and L E Bermudez. The *Mycobacterium avium* subsp. *paratuberculosis* 35 kDa protein plays a role in invasion of bovine epithelial cells. *Microbiology*, 149(Pt 8): 2061–9, 2003.

A F Barakat, A Hegazy, R E Farag, A A Baky, L F Arafa, and A Farouk. Role of Interferon-gamma and immune response biomarkers in predicting IFN-alpha responsiveness and treatment outcome in patients with hepatitis C virus. *Int J Virol*, 8(4):288–98, 2012.

T Barrett, T O Suzek, D B Troup, S E Wilhite, W-C Ngau, P Ledoux, D Rudnev, A E Lash, W Fujibuchi, and R Edgar. NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res*, 33(Database issue):D562–6, 2005.

T Barrett, D B Troup, S E Wilhite, P Ledoux, D Rudnev, C Evangelista, I F Kim, A Soboleva, M Tomashevsky, and R Edgar. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35(Database issue):D760–5, 2007.

T Barrett, D B Troup, S E Wilhite, P Ledoux, D Rudnev, C Evangelista, I F Kim, A Soboleva, M Tomashevsky, K A Marshall, K H Phillippy, P M Sherman, R N Muertter, and R Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue):D885–90, 2009.

T Barrett, D B Troup, S E Wilhite, P Ledoux, C Evangelista, I F Kim, M Tomashevsky, K A Marshall, K H Phillippy, P M Sherman, R N Muertter, M Holko, O Ayanbule, A Yefanov, and A Soboleva. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–10, 2011.

T Barrett, S E Wilhite, P Ledoux, C Evangelista, I F Kim, M Tomashevsky, K A Marshall, K H Phillippy, P M Sherman, M Holko, A Yefanov, H Lee, N Zhang, C L Robertson, N Serova, S Davis, and A Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41(Database issue): D991–5, 2013.

A Barsky, J L Gardy, R E W Hancock, and T Munzner. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, 23(8):1040–2, 2007.

S Basu and D Plewczynski. AMS 3.0: prediction of post-translational modifications. *BMC Bioinformatics*, 11:210, 2010.

I Ben-Porath, M W Thomson, V J Carey, R Ge, G W Bell, A Regev, and R A Weinberg. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet*, 40 (5):499–507, 2008.

Y Benjamini and Y Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B*, 57:289–300, 1995.

E Berezikov, V Guryev, R H A Plasterk, and E Cuppen. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res*, 14(1):170–8, 2004.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42, 2000.

E A Berry, A R Dalby, and Z R Yang. Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. *Comput Biol Chem*, 28(1):75–85, 2004.

C M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

A K Biswas, N Noman, and A R Sikder. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics*, 11:273, 2010.

N Blom, A Kreegipuu, and S Brunak. PhosphoBase: a database of phosphorylation sites. *Nucleic Acids Res*, 26(1):382–6, 1998.

N Blom, S Gammeltoft, and S Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 294(5):1351–62, 1999.

N Blom, T Sicheritz-Pontén, R Gupta, S Gammeltoft, and S Brunak. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, 4(6):1633–49, 2004.

T W Bodnarchuk, S Napper, N Rapin, and V Misra. Mechanism for the induction of cell death in ONS-76 medulloblastoma cells by Zhangfei/CREB-ZF. *J Neurooncol*, 109(3):485–501, 2012.

P J Boersema, S Mohammed, and A J R Heck. Phosphopeptide fragmentation and analysis by mass spectrometry. *J Mass Spectrom*, 44(6):861–78, 2009.

J S Booth, R Arsenault, S Napper, P J Griebel, A A Potter, L A Babiuk, and G K Mutwiri. TLR9 signaling failure renders Peyer's patch regulatory B cells unresponsive to stimulation with CpG oligodeoxynucleotides. *J Innate Immun*, 2(5):483–94, 2010.

E Boutet, D Lieberherr, M Tognolli, M Schneider, and A Bairoch. UniProtKB/Swiss-Prot. *Methods Mol Biol*, 406:89–112, 2007.

A Brazma, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, W Ansorge, C A Ball, H C Causton, T Gaasterland, P Glenisson, F C Holstege, I F Kim, V Markowitz, J C Matese, H Parkinson, A Robinson, U Sarkans, S Schulze-Kremer, J Stewart, R Taylor, J Vilo, and M Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4):365–71, 2001.

A Brazma, H Parkinson, U Sarkans, M Shojatalab, J Vilo, N Abeygunawardena, E Holloway, M Kapushesky, P Kemmeren, G G Lara, A Oezcimen, P Rocca-Serra, and S-A Sansone. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 31(1):68–71, 2003.

L Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

K Breuer, A K Foroushani, M R Laird, C Chen, A Sribnaia, R Lo, G L Winsor, R E W Hancock, F S L Brinkman, and D J Lynn. InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res*, 41(Database issue):D1228–33, 2013.

R I Brinkworth, R A Breinl, and B Kobe. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc Natl Acad Sci U S A*, 100(1):74–9, 2003.

Y-H Bu, Y-L He, H-D Zhou, W Liu, D Peng, A-G Tang, L-L Tang, H Xie, Q-X Huang, X-H Luo, and E-Y Liao. Insulin receptor substrate 1 regulates the cellular differentiation and the matrix metallopeptidase expression of preosteoblastic cells. *J Endocrinol*, 206(3):271–7, 2010.

D R Caffrey, S Somaroo, J D Hughes, J Mintseris, and E S Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 13(1):190–202, 2004.

N W Calderone. Insect pollinated crops, insect pollinators and US agriculture: trend analysis of aggregate data for the period 1992-2009. *PLoS One*, 7(5):e37235, 2012.

R Caspi, H Foerster, C A Fulcher, P Kaipa, M Krummenacker, M Latendresse, S Paley, S Y Rhee, A G Shearer, C Tissier, T C Walk, P Zhang, and P D Karp. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*, 36(Database issue):D623–31, 2008.

A Champion, M Kreis, K Mockaitis, A Picaud, and Y Henry. *Arabidopsis* kinome: after the casting. *Funct Integr Genomics*, 4(3):163–87, 2004.

L Chang and M Karin. Mammalian MAP kinase signalling cascades. *Nature*, 410(6824):37–40, 2001.

C Charavaryamath, P Fries, S Gomis, C Bell, K Doig, L L Guan, A Potter, S Napper, and P J Griebel. Mucosal changes in a long-term bovine intestinal segment model following removal of ingesta and microflora. *Gut Microbes*, 2(3):134–44, 2011.

C Charavaryamath, P Gonzalez-Cano, P Fries, S Gomis, K Doig, E Scruten, A Potter, S Napper, and P J Griebel. Host responses to persistent *Mycobacterium avium* subspecies *paratuberculosis* infection in surgically isolated bovine ileal segments. *Clin Vaccine Immunol*, 20(2):156–65, 2013.

A L Chernorudskiy, A Garcia, E V Eremin, A S Shorina, E V Kondratieva, and M R Gainullin. UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics*, 8:126, 2007.

J M Cherry, C Adler, C Ball, S A Chervitz, S S Dwight, E T Hester, Y Jia, G Juvik, T Roe, M Schroeder, S Weng, and D Botstein. SGD: Saccharomyces Genome Database. *Nucleic Acids Res*, 26(1):73–9, 1998.

J M Cherry, E L Hong, C Amundsen, R Balakrishnan, G Binkley, E T Chan, K R Christie, M C Costanzo, S S Dwight, S R Engel, D G Fisk, J E Hirschman, B C Hitz, K Karra, C J Krieger, S R Miyasato, R S Nash, J Park, M S Skrzypek, M Simison, S Weng, and E D Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*, 40(Database issue):D700–5, 2012.

A Chiang and R P Million. Personalized medicine in oncology: next generation. *Nat Rev Drug Discov*, 10 (12):895–6, 2011.

A C Christensen. Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. *Genome Biol Evol*, 5(6):1079–86, 2013.

P Cohen. Protein kinases–the major drug targets of the twenty-first century? *Nat Rev Drug Discov*, 1(4): 309–15, 2002.

J R Cole, Q Wang, J A Fish, B Chai, D M McGarrell, Y Sun, C T Brown, A Porras-Alfaro, C R Kuske, and J M Tiedje. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*, 42(1):D633–42, 2014.

D Conway and J M White. *Machine Learning for Hackers*. O'Reilly Media, 2012.

C Cortes and V Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–97, 1995.

P M Coussens. Model for immune responses to *Mycobacterium avium* subspecies *paratuberculosis* in cattle. *Infect Immun*, 72(6):3089–96, 2004.

D L Cox-Foster, S Conlan, E C Holmes, G Palacios, J D Evans, N A Moran, P-L Quan, T Briese, M Hornig, D M Geiser, V Martinson, D vanEngelsdorp, A L Kalkstein, A Drysdale, J Hui, J Zhai, L Cui, S K Hutchison, J F Simons, M Egholm, J S Pettis, and W I Lipkin. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, 318(5848):283–7, 2007.

D Croft, G O'Kelly, G Wu, R Haw, M Gillespie, L Matthews, M Caudy, P Garapati, G Gopinath, B Jassal, S Jupe, I Kalatskaya, S Mahajan, B May, N Ndegwa, E Schmidt, V Shamovsky, C Yung, E Birney, H Hermjakob, P D'Eustachio, and L Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*, 39(Database issue):D691–7, 2011.

X Cui and G A Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210, 2003.

T H Dang, K Van Leemput, A Verschoren, and K Laukens. Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics*, 24(24):2857–64, 2008.

J E Darnell, Jr. STATs and gene regulation. *Science*, 277(5332):1630–5, 1997.

M H de Borst, S H Diks, J Bolbrinker, M W Schellings, M B A van Dalen, M P Peppelenbosch, R Kreutz, Y M Pinto, G Navis, and H van Goor. Profiling of the renal kinome: a novel tool to identify protein kinases involved in angiotensin II-dependent hypertensive renal damage. *Am J Physiol Renal Physiol*, 293 (1):F428–37, 2007.

T Z DeSantis, P Hugenholtz, N Larsen, M Rojas, E L Brodie, K Keller, T Huber, D Dalevi, P Hu, and G L Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, 72(7):5069–72, 2006.

G Di Prisco, F Pennacchio, E Caprio, H F Boncristiani, Jr, J D Evans, and Y Chen. *Varroa destructor* is an effective vector of Israeli acute paralysis virus in the honeybee, *Apis mellifera*. *J Gen Virol*, 92(Pt 1): 151–5, 2011.

F Diella, S Cameron, C Gemünd, R Linding, A Via, B Kuster, T Sicheritz-Pontén, N Blom, and T J Gibson. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5:79, 2004.

F Diella, C M Gould, C Chica, A Via, and T J Gibson. Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res*, 36(Database issue):D240–4, 2008.

V Dietemann, F Nazzi, S J Martin, D L Anderson, B Locke, K S Delaplane, Q Wauquiez, C Tannahill, E Frey, B Ziegelmann, P Rosenkranz, and J D Ellis. Standard methods for varroa research. *J Apicult Res*, 52(1), 2013.

S H Diks, K Kok, T O'Toole, D W Hommes, P van Dijken, J Joore, and M P Peppelenbosch. Kinome profiling for studying lipopolysaccharide signal transduction in human peripheral blood mononuclear cells. *J Biol Chem*, 279(47):49206–13, 2004.

S H Diks, K Parikh, M van der Sijde, J Joore, T Ritsema, and M P Peppelenbosch. Evidence for a minimal eukaryotic phosphoproteome? *PLoS One*, 2(1):e777, 2007.

H Dinkel, C Chica, A Via, C M Gould, L J Jensen, T J Gibson, and F Diella. Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res*, 39(Database issue):D261–7, 2011.

R Döffinger, E Jouanguy, S Dupuis, M C Fondanèche, J L Stephan, J F Emile, S Lamhamedi-Cherradi, F Altare, A Pallier, G Barcenas-Morales, E Meinl, C Krause, S Pestka, R D Schreiber, F Novelli, and J L Casanova. Partial interferon-gamma receptor signaling chain deficiency in a patient with bacille Calmette-Guérin and *Mycobacterium abscessus* infection. *J Infect Dis*, 181(1):379–84, 2000.

S E Dorman and S M Holland. Mutation in the signal-transducing chain of the interferon-gamma receptor and susceptibility to mycobacterial infection. *J Clin Invest*, 101(11):2364–9, 1998.

S Draghici. *Data analysis tools for DNA microarrays*. Chapman & Hall/CRC, 2003.

Y Du, N Xu, M Lu, and T Li. hUbiquitome: a database of experimentally verified ubiquitination cascades in humans. *Database (Oxford)*, 2011:bar055, 2011.

S Dudoit, P J Shaffer, and J C Boldrick. Multiple hypothesis testing in microarray experiments. *Stat Science*, 18:71–103, 2003.

M J Duffy. Carcinoembryonic antigen as a marker for colorectal cancer: is it clinically useful? *Clin Chem*, 47(4):624–30, 2001.

A K Dunker, C J Brown, J D Lawson, L M Iakoucheva, and Z Obradović. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–82, 2002.

B P Durbin, J S Hardin, D M Hawkins, and D M Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl 1:S105–10, 2002.

P Durek, C Schudoma, W Weckwerth, J Selbig, and D Walther. Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics*, 10:117, 2009.

P Durek, R Schmidt, J L Heazlewood, A Jones, D MacLean, A Nagel, B Kersten, and W X Schulze. PhosPhAt: the *Arabidopsis thaliana* phosphorylation site database. An update. *Nucleic Acids Res*, 38 (Database issue):D828–34, 2010.

V I Dyukova, N V Shilova, O E Galanina, A Yu Rubina, and N V Bovin. Design of carbohydrate multiarrays. *Biochim Biophys Acta*, 1760(4):603–9, 2006.

A D Ebert, J Yu, F F Rose, Jr, V B Mattis, C L Lorson, J A Thomson, and C N Svendsen. Induced pluripotent stem cells from a spinal muscular atrophy patient. *Nature*, 457(7227):277–80, 2009.

R Edgar, M Domrachev, and A E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, 2002.

R Eglen and T Reisine. Drug discovery and the human kinome: recent trends. *Pharmacol Ther*, 130(2): 144–56, 2011.

M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.

J L E Ellingson, J L Anderson, J J Koziczkowski, R P Radcliff, S J Sloan, S E Allen, and N M Sullivan. Detection of viable *Mycobacterium avium* subsp. *paratuberculosis* in retail pasteurized whole milk by two culture methods and PCR. *J Food Prot*, 68(5):966–72, 2005.

ENCODE Project Consortium, B E Bernstein, E Birney, I Dunham, E D Green, C Gunter, and M Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

S R Engel, R Balakrishnan, G Binkley, K R Christie, M C Costanzo, S S Dwight, D G Fisk, J E Hirschman, B C Hitz, E L Hong, C J Krieger, M S Livstone, S R Miyasato, R Nash, R Oughtred, J Park, M S Skrzypek, S Weng, E D Wong, K Dolinski, D Botstein, and J M Cherry. Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res*, 38(Database issue):D433–6, 2010.

J D Evans and R S Schwarz. Bees brought to their knees: microbes affecting honey bee health. *Trends Microbiol*, 19(12):614–20, 2011.

B Everitt. *Cluster Analysis*. Heinemann Educ., 1974.

H-C Fan, X Zhang, and P A McNaughton. Activation of the TRPV4 ion channel is enhanced by phosphorylation. *J Biol Chem*, 284(41):27884–91, 2009.

A Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat Methods Med Res*, 17(4):347–88, 2007.

T Fawcett. An introduction to ROC analysis. *Pattern Recogn Lett*, 27(8):861–74, 2006.

J Felenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–91, 1985.

X M Fernández-Suárez and M Y Galperin. The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Res*, 41(Database issue):D1–7, 2013.

S B Ficarro, M L McCleland, P T Stukenberg, D J Burke, M M Ross, J Shabanowitz, D F Hunt, and F M White. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat Biotechnol*, 20(3):301–5, 2002.

G Finak, S Sadekova, F Pepin, M Hallett, S Meterissian, F Halwani, K Khetani, M Souleimanova, B Zabolotny, A Omeroglu, and M Park. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res*, 8(5):R58, 2006.

H A Fletcher, A Keyser, M Bowmaker, P C Sayles, G Kaplan, G Hussey, A V S Hill, and W A Hanekom. Transcriptional profiling of mycobacterial antigen-induced responses in infants vaccinated with BCG at birth. *BMC Med Genomics*, 2:10, 2009.

E Frank, M Hall, L Trigg, G Holmes, and I H Witten. Data mining in bioinformatics using Weka. *Bioinformatics*, 20(15):2479–81, 2004.

R C Friedman, K K-H Farh, C B Burge, and D P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, 2009.

Functional Genomics Data Society. MGED—Workgroups—MIAME—Journals, 2010. URL `http://www.mged.org/Workgroups/MIAME/journals.html`.

K Fundel, J Haag, P M Gebhard, R Zimmer, and T Aigner. Normalization strategies for mRNA expression data in cartilage research. *Osteoarthritis Cartilage*, 16(8):947–55, 2008a.

K Fundel, R Küffner, T Aigner, and R Zimmer. Normalization and gene p-value estimation: issues in microarray data processing. *Bioinform Biol Insights*, 2:291–305, 2008b.

J Gao and D Xu. The Musite open-source framework for phosphorylation-site prediction. *BMC Bioinformatics*, 11 Suppl 12:S9, 2010.

J Gao, G K Agrawal, J J Thelen, Z Obradovic, A K Dunker, and D Xu. A new machine learning approach for protein phosphorylation site prediction in plants. *Lect Notes Comput Sci*, 5462/2009:18–29, 2009a.

J Gao, G K Agrawal, J J Thelen, and D Xu. P3DB: a plant protein phosphorylation database. *Nucleic Acids Res*, 37(Database issue):D960–2, 2009b.

J Gao, J J Thelen, A K Dunker, and D Xu. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics*, 9(12):2586–600, 2010.

S B Giese and P Ahrens. Detection of *Mycobacterium avium* subsp. *paratuberculosis* in milk from clinically affected cows by PCR and culture. *Vet Microbiol*, 77(3-4):291–7, 2000.

K Ginalski, A Elofsson, D Fischer, and L Rychlewski. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, 19(8):1015–8, 2003.

E Glaab, J M Garibaldi, and N Krasnogor. vrmlgen: an R Package for 3D data visualization on the web. *Journal of Statistical Software*, 36(8):1–18, 2010.

F Gnad, S Ren, J Cox, J V Olsen, B Macek, M Oroshi, and M Mann. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol*, 8(11):R250, 2007.

F Gnad, J Gunawardena, and M Mann. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res*, 39(Database issue):D253–60, 2011.

J Goecks, A Nekrutenko, J Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8): R86, 2010.

D Gonzalez de Castro, P A Clarke, B Al-Lazikani, and P Workman. Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. *Clin Pharmacol Ther*, 93(3):252–9, 2013.

D M Goodstein, S Shu, R Howson, R Neupane, R D Hayes, J Fazo, T Mitros, W Dirks, U Hellsten, N Putnam, and D S Rokhsar. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, 40 (Database issue):D1178–86, 2012.

T D Gould and H K Manji. Glycogen synthase kinase-3: a putative molecular target for lithium mimetic drugs. *Neuropsychopharmacology*, 30(7):1223–37, 2005.

L M Graves, J S Duncan, M C Whittle, and G L Johnson. The dynamic nature of the kinome. *Biochem J*, 450(1):1–8, 2013.

Sarah S Greenleaf and Claire Kremen. Wild bees enhance honey bees' pollination of hybrid sunflower. *Proc Natl Acad Sci U S A*, 103(37):13890–5, 2006.

Pamela G Gregory, Jay D Evans, Thomas Rinderer, and Lilia de Guzman. Conditional immune-gene suppression of honeybees parasitized by Varroa mites. *J Insect Sci*, 5:7, 2005.

M A M Groenen et al. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491(7424):393–8, 2012.

G K Gronvall, D Trent, L Borio, R Brey, L Nagao, and Alliance for Biosecurity. The FDA animal efficacy rule and biodefense. *Nat Biotechnol*, 25(10):1084–7, 2007.

M Grzmil, P Morin, Jr, M M Lino, A Merlo, S Frank, Y Wang, G Moncayo, and B A Hemmings. MAP kinase-interacting kinase 1 regulates SMAD2-dependent TGF-β signaling pathway in human glioblastoma. *Cancer Res*, 71(6):2392–402, 2011.

Y Gu, J Rosenblatt, and D O Morgan. Cell cycle regulation of CDK2 activity by phosphorylation of Thr160 and Tyr15. *EMBO J*, 11(11):3995–4005, 1992.

R Gupta, H Birch, K Rapacki, S Brunak, and J E Hansen. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res*, 27(1):370–2, 1999.

S K Hanks and T Hunter. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J*, 9(8):576–96, 1995.

J E Hansen, O Lund, J Nilsson, K Rapacki, and S Brunak. O-GLYCBASE Version 3.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res*, 26(1):387–9, 1998.

J R Harbo and J W Harris. Responses to Varroa by honey bees with different levels of Varroa Sensitive Hygiene. *J Apic Res*, 48:156–161, 2009.

H C Harsha and A Pandey. Phosphoproteomics in cancer. *Mol Oncol*, 4(6):482–95, 2010.

J A Hartigan. *Clustering Algorithms*. Wiley, 1975.

A L Hazen, S H Diks, J A Wahle, G M Fuhler, M P Peppelenbosch, and W G Kerr. Major remodelling of the murine stem cell kinome following differentiation in the hematopoietic compartment. *J Proteome Res*, 10(8):3542–50, 2011.

J L Heazlewood, P Durek, J Hummel, J Selbig, W Weckwerth, D Walther, and W X Schulze. PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res*, 36(Database issue):D1015–21, 2008.

W R Hein and P J Griebel. A road less travelled: large animal models in immunological research. *Nat Rev Immunol*, 3(1):79–84, 2003.

S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9, 1992.

A L K Hestvik, Z Hmama, and Y Av-Gay. Kinome analysis of host response to mycobacterial infection: a novel technique in proteomics. *Infect Immun*, 71(10):5514–22, 2003.

M A T Hildebrandt, W Tan, P Tamboli, M Huang, Y Ye, J Lin, J-S Lee, C G Wood, and X Wu. Kinome expression profiling identifies IKBKE as a predictor of overall survival in clear cell renal cell carcinoma patients. *Carcinogenesis*, 33(4):799–803, 2012.

G E Hinton and S T Roweis. *Stochastic Neighbor Embedding*, volume 15, pages 833–840. MIT Press, Cambridge, MA, USA, 2002.

M Hjerrild, A Stensballe, T E Rasmussen, C B Kofoed, N Blom, T Sicheritz-Ponten, M R Larsen, S Brunak, O N Jensen, and S Gammeltoft. Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *J Proteome Res*, 3(3):426–33, 2004.

Majbrit Hjerrild and Steen Gammeltoft. Phosphoproteomics toolbox: computational biology, protein chemistry and mass spectrometry. *FEBS Lett*, 580(20):4764–70, 2006.

Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, 443(7114):931–49, 2006.

J J M Hoozemans, R Hilhorst, R Ruijtenbeek, A J M Rozemuller, and S M van der Vies. Protein kinase activity profiling of postmortem human brain tissue. *Neurodegener Dis*, 10(1-4):46–8, 2012.

J E Hopcroft, R Motwani, and J D Ullman. *Introduction to Automata Theory, Languages, and Computation.* Prentice Hall, 3rd edition, 2006.

A L Hopkins and C R Groom. The druggable genome. *Nat Rev Drug Discov*, 1(9):727–30, 2002.

P V Hornbeck, I Chabra, J M Kornhauser, E Skrzypek, and B Zhang. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–61, 2004.

P V Hornbeck, J M Kornhauser, S Tkachev, B Zhang, E Skrzypek, B Murray, V Latham, and M Sullivan. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res*, 40(Database issue): D261–70, 2012.

B T Houseman and M Mrksich. Towards quantitative assays with peptide chips: a surface engineering approach. *Trends Biotechnol*, 20(7):279–81, 2002.

B T Houseman, J H Huh, S J Kron, and M Mrksich. Peptide chips for the quantitative evaluation of protein kinase activity. *Nat Biotechnol*, 20(3):270–4, 2002.

H-D Huang, T-Y Lee, S-W Tzeng, and J-T Horng. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res*, 33(Web Server issue):W226–9, 2005a.

H-D Huang, T-Y Lee, S-W Tzeng, L-C Wu, J-T Horng, A-P Tsou, and K-T Huang. Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J Comput Chem*, 26(10): 1032–41, 2005b.

W Huber, A von Heydebreck, H Sültmann, A Poustka, and M Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

W Huber, A von Heydebreck, H Sueltmann, A Poustka, and M Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol*, 2:Article3, 2003.

T Hunter and G D Plowman. The protein kinases of budding yeast: six score and more. *Trends Biochem Sci*, 22(1):18–22, 1997.

D H Huson and C Xie. A poor man's BLASTX–high-throughput metagenomic protein database search using PAUDA. *Bioinformatics*, 30(1):38–9, 2014.

E L Huttlin, M P Jedrychowski, J E Elias, T Goswami, R Rad, S A Beausoleil, J Villén, W Haas, M E Sowa, and S P Gygi. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143 (7):1174–89, 2010.

L M Iakoucheva, P Radivojac, C J Brown, T R O'Connor, J G Sikes, Z Obradovic, and A K Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, 32(3):1037–49, 2004.

Ingenuity Systems. Ingenuity Pathway Analysis. http://www.ingenuity.com/products/ipa, 2013.

C R Ingrell, M L Miller, O N Jensen, and N Blom. NetPhosYeast: prediction of protein phosphorylation sites in yeast. *Bioinformatics*, 23(7):895–7, 2007.

S Jalal, J Kindrachuk, and S Napper. Phosphoproteome and kinome analysis: unique perspectives on the same problem. *Curr Anal Chem*, 3(1):1–15, 2007.

S Jalal, R Arsenault, A A Potter, L A Babiuk, P J Griebel, and S Napper. Genome to kinome: species-specific peptide arrays for kinome analysis. *Sci Signal*, 2(54):pl1, 2009.

N Japkowicz and S Stephen. The class imbalance problem: a systematic study. *Intelligent Data Analysis*, 6 (5), 2002.

J L Jiménez, B Hegemann, J R A Hutchins, J-M Peters, and R Durbin. A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol*, 8(5): R90, 2007.

T Joachims. *Advances in Kernel Methods: Support Vector Learning*, chapter 11. MIT Press, 1998.

S A Johnson and T Hunter. Kinomics: methods for deciphering the kinome. *Nat Methods*, 2(1):17–25, 2005.

G Joshi-Tope, M Gillespie, I Vastrik, P D'Eustachio, E Schmidt, B de Bono, B Jassal, G R Gopinath, G R Wu, L Matthews, S Lewis, E Birney, and L Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database issue):D428–32, 2005.

I Jung, A Matsuyama, M Yoshida, and D Kim. PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. *BMC Bioinformatics*, 11 Suppl 1:S10, 2010.

K Kandasamy, S S Mohan, R Raju, S Keerthikumar, G S S Kumar, A K Venugopal, D Telikicherla, J D Navarro, S Mathivanan, C Pecquet, S K Gollapudi, S G Tattikota, S Mohan, H Padhukasahasram, Y Subbannayya, R Goel, H K C Jacob, J Zhong, R Sekhar, V Nanjappa, L Balakrishnan, R Subbaiah, Y L Ramachandra, B A Rahiman, T S K Prasad, J-X Lin, J C D Houtman, S Desiderio, J-C Renauld, S N Constantinescu, O Ohara, T Hirano, M Kubo, S Singh, P Khatri, S Draghici, G D Bader, C Sander, W J Leonard, and A Pandey. NetPath: a public resource of curated signal transduction pathways. *Genome Biol*, 11(1):R3, 2010.

M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

M Kanehisa, S Goto, M Hattori, K F Aoki-Kinoshita, M Itoh, S Kawashima, T Katayama, M Araki, and M Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34 (Database issue):D354–7, 2006.

M Kanehisa, S Goto, M Furumichi, M Tanabe, and M Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):D355–60, 2010.

O Karpenko, L Huang, and Y Dai. A probabilistic meta-predictor for the MHC class II binding peptides. *Immunogenetics*, 60(1):25–36, 2008.

L Kaufman and P J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

S Kawashima, P Pokarowski, M Pokarowska, A Kolinski, T Katayama, and M Kanehisa. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202–5, 2008.

B E Kemp, D J Graves, E Benjamini, and E G Krebs. Role of multiple basic residues in determining the substrate specificity of cyclic AMP-dependent protein kinase. *J Biol Chem*, 252(14):4888–94, 1977.

M K Kerr, M Martin, and G A Churchill. Analysis of variance for gene expression microarray data. *J Comput Biol*, 7(6):819–37, 2000.

S Khare, S D Lawhon, K L Drake, J E S Nunes, J F Figueiredo, C A Rossetti, T Gull, R E Everts, H A Lewin, C L Galindo, H R Garner, and L G Adams. Systems biology analysis of gene expression during in vivo *Mycobacterium avium paratuberculosis* enteric colonization reveals role for immune tolerance. *PLoS One*, 7(8):e42127, 2012.

S Kilpinen, K Ojala, and O Kallioniemi. Analysis of kinase gene expression patterns across 5681 human tissue samples reveals functional genomic taxonomy of the kinome. *PLoS One*, 5(12):e15068, 2010.

J H Kim, J Lee, B Oh, K Kimm, and I Koh. Prediction of phosphorylation sites using SVMs. *Bioinformatics*, 20(17):3179–84, 2004.

S-H Kim and C-E Lee. Counter-regulation mechanism of IL-4 and IFN-α signal transduction through cytosolic retention of the pY-STAT6:pY-STAT2:p48 complex. *Eur J Immunol*, 41(2):461–72, 2011.

J Kindrachuk, E Scruten, S Attah-Poku, K Bell, A Potter, L A Babiuk, P J Griebel, and S Napper. Stability, toxicity, and biological activity of host defense peptide BMAP28 and its inversed and retro-inversed isomers. *Biopolymers*, 96(1):14–24, 2011.

J Kindrachuk, R Arsenault, A Kusalik, K N Kindrachuk, B Trost, S Napper, P B Jahrling, and J E Blaney. Systems kinomics demonstrates Congo Basin monkeypox virus infection selectively modulates host cell signaling responses as compared to West African monkeypox virus. *Mol Cell Proteomics*, 11(6):M111.015701, 2012.

J Kitchen, R E Saunders, and J Warwicker. Charge environments around phosphorylation sites in proteins. *BMC Struct Biol*, 8:19, 2008.

A-M Klein, B E Vaissière, J H Cane, I Steffan-Dewenter, S A Cunningham, C Kremen, and T Tscharntke. Importance of pollinators in changing landscapes for world crops. *Proc Biol Sci*, 274(1608):303–13, 2007.

J Kleinnijenhuis, M Oosting, L A B Joosten, M G Netea, and R Van Crevel. Innate immune recognition of *Mycobacterium tuberculosis*. *Clin Dev Immunol*, 2011:405310, 2011.

J D R Knight, T Pawson, and A-C Gingras. Profiling the kinome: current capabilities and future challenges. *J Proteomics*, 81:43–55, 2013.

B Kobe, T Kampmann, J K Forwood, P Listwan, and R I Brinkworth. Substrate specificity of protein kinases and computational prediction of substrates. *Biochim Biophys Acta*, 1754(1-2):200–9, 2005.

M Koenig and N Grabe. Highly specific prediction of phosphorylation sites in proteins. *Bioinformatics*, 20 (18):3620–7, 2004.

A P Koets, G Adugna, L L Janss, H J van Weering, C H Kalis, G H Wentink, V P Rutten, and Y H Schukken. Genetic variation of susceptibility to *Mycobacterium avium* subsp. *paratuberculosis* infection in dairy cattle. *J Dairy Sci*, 83(11):2702–8, 2000.

T Kohonen. The self-organizing map. *Proc IEEE*, 78(9):1464–80, 1990.

A Kreegipuu, N Blom, and S Brunak. PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res*, 27(1):237–9, 1999.

P Lamesch, T Z Berardini, D Li, D Swarbreck, C Wilks, R Sasidharan, R Muller, K Dreher, D L Alexander, M Garcia-Hernandez, A S Karthikeyan, C H Lee, W D Nelson, L Ploetz, S Singh, A Wensel, and E Huala. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*, 40(Database issue):D1202–10, 2012.

E S Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

J Le, J X Lin, D Henriksen-DeStefano, and J Vilcek. Bacterial lipopolysaccharide-induced interferon-gamma production: roles of interleukin 1 and interleukin 2. *J Immunol*, 136(12):4525–30, 1986.

Y Le Conte, G de Vaublanc, D Crauser, F Jeanne, J-C Rousselle, and J M Bécard. Honey bee colonies that have survived *Varroa destructor*. *Apidologie*, 38:566–572, 2007.

Y Le Conte, C Alaux, J-F Martin, J R Harbo, J W Harris, C Dantec, D Séverac, S Cros-Arteil, and M Navajas. Social immunity in honeybees (*Apis mellifera*): transcriptome analysis of varroa-hygienic behaviour. *Insect Mol Biol*, 20(3):399–408, 2011.

T-Y Lee, N A Bretaña, and C-T Lu. PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity. *BMC Bioinformatics*, 12:261, 2011.

J A Lees, K J Buchkovich, D R Marshak, C W Anderson, and E Harlow. The retinoblastoma protein is phosphorylated on multiple sites by human cdc2. *EMBO J*, 10(13):4279–90, 1991.

L Li, C Wu, H Huang, K Zhang, J Gan, and S S-C Li. Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic Acids Res*, 36(10):3263–73, 2008a.

T Li, F Li, and X Zhang. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins*, 70(2):404–14, 2008b.

T Li, P Du, and N Xu. Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. *PLoS One*, 5(11):e15411, 2010.

W Li and A Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–9, 2006.

Y Li, R J Arsenault, B Trost, J Slind, P J Griebel, S Napper, and A Kusalik. A systematic approach for analysis of peptide array kinome data. *Sci Signal*, 5(220):pl2, 2012.

I Lian, J Kim, H Okazawa, J Zhao, B Zhao, J Yu, A Chinnaiyan, M A Israel, L S B Goldstein, R Abujarour, S Ding, and K-L Guan. The role of YAP transcription coactivator in regulating stem cell self-renewal and differentiation. *Genes Dev*, 24(11):1106–18, 2010.

H Lilja, D Ulmert, and A J Vickers. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer*, 8(4):268–78, 2008.

R Linding, L J Jensen, G J Ostheimer, M A T M van Vugt, C Jørgensen, I M Miron, F Diella, K Colwill, L Taylor, K Elder, P Metalnikov, V Nguyen, A Pasculescu, J Jin, J G Park, L D Samson, J R Woodgett, R B Russell, P Bork, M B Yaffe, and T Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–26, 2007.

J Liu, S Kang, C Tang, L B M Ellis, and T Li. Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acids Res*, 35(15):e96, 2007.

Z Liu, J Cao, X Gao, Y Zhou, L Wen, X Yang, X Yao, J Ren, and Y Xue. CPLA 1.0: an integrated database of protein lysine acetylation. *Nucleic Acids Res*, 39(Database issue):D1029–34, 2011.

M Lodesani and C Costa. Limits of chemotherapy in beekeeping: development of resistance and the problem of residues. *Bee World*, 86:102–9, 2005.

M Löwenberg, J Tuynman, J Bilderbeek, T Gaber, F Buttgereit, S van Deventer, M Peppelenbosch, and D Hommes. Rapid immunosuppressive effects of glucocorticoids mediated through Lck and Fyn. *Blood*, 106(5):1703–10, 2005.

M Löwenberg, J Tuynman, M Scheffer, A Verhaar, L Vermeulen, S van Deventer, D Hommes, and M Peppelenbosch. Kinome analysis reveals nongenomic glucocorticoid receptor-dependent inhibition of insulin signaling. *Endocrinology*, 147(7):3555–62, 2006.

J A Ludwig and J N Weinstein. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer*, 5(11):845–56, 2005.

A Lueking, M Horn, H Eickhoff, K Büssow, H Lehrach, and G Walter. Protein microarrays for gene expression and antibody screening. *Anal Biochem*, 270(1):103–11, 1999.

A V Lukashin and M Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26 (4):1107–15, 1998.

M Lukk, M Kapushesky, J Nikkilä, H Parkinson, A Goncalves, W Huber, E Ukkonen, and A Brazma. A global map of human gene expression. *Nat Biotechnol*, 28(4):322–4, 2010.

D J Lynn, G L Winsor, C Chan, N Richard, M R Laird, A Barsky, J L Gardy, F M Roche, T H W Chan, N Shah, R Lo, M Naseer, J Que, M Yau, M Acab, D Tulpan, M D Whiteside, A Chikatamarla, B Mah, T Munzner, K Hokamp, R E W Hancock, and F S L Brinkman. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol Syst Biol*, 4:218, 2008.

F J Maathuis. Conservation of protein phosphorylation sites within gene families and across species. *Plant Signal Behav*, 3(11):1011–3, 2008.

P Määttänen, B Trost, E Scruten, A Potter, A Kusalik, P Griebel, and S Napper. Divergent immune responses to *Mycobacterium avium* subsp. *paratuberculosis* infection correlate with kinome responses at the site of intestinal infection. *Infect Immun*, 81(8):2861–72, 2013.

J A MacDonald, A J Mackey, W R Pearson, and T A J Haystead. A strategy for the rapid identification of phosphorylation sites in the phosphoproteome. *Mol Cell Proteomics*, 1(4):314–22, 2002.

B Macek, I Mijakovic, J V Olsen, F Gnad, C Kumar, P R Jensen, and M Mann. The serine/threonine/tyrosine phosphoproteome of the model bacterium *Bacillus subtilis*. *Mol Cell Proteomics*, 6(4):697–707, 2007.

B Macek, F Gnad, B Soufi, C Kumar, J V Olsen, I Mijakovic, and M Mann. Phosphoproteome analysis of *E. coli* reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics*, 7(2):299–307, 2008.

A Maddigan, L Truitt, R Arsenault, T Freywald, O Allonby, J Dean, A Narendran, J Xiang, A Weng, S Napper, and A Freywald. EphB receptors trigger Akt activation and suppress Fas receptor-induced apoptosis in malignant T lymphocytes. *J Immunol*, 187(11):5983–94, 2011.

B L Maidak, G J Olsen, N Larsen, R Overbeek, M J McCaughey, and C R Woese. The RDP (Ribosomal Database Project). *Nucleic Acids Res*, 25(1):109–11, 1997.

M Mann, S E Ong, M Grønborg, H Steen, O N Jensen, and A Pandey. Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol*, 20(6):261–8, 2002.

G Manning, D B Whyte, R Martinez, T Hunter, and S Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–34, 2002.

C Marcé, P Ezanno, H Seegers, D U Pfeiffer, and C Fourichon. Within-herd contact structure and transmission of *Mycobacterium avium* subspecies *paratuberculosis* in a persistently infected dairy cattle herd. *Prev Vet Med*, 100(2):116–25, 2011.

K V Mardia, J T Kent, and J M Bibby. *Multivariate Analysis*. Academic Press, 1979.

S J Martin, A C Highfield, L Brettell, E M Villalobos, G E Budge, M Powell, S Nikaido, and D C Schroeder. Global honey bee viral landscape altered by a parasitic mite. *Science*, 336(6086):1304–6, 2012.

S J Martin, B V Ball, and N L Carreck. The role of deformed wing virus in the initial collapse of varroa infested honey bee colonies in the UK. *J Apicult Res*, 52(5):251–8, 2013.

I Matic, B Macek, M Hilger, T C Walther, and M Mann. Phosphorylation of SUMO-1 occurs *in vivo* and is conserved through evolution. *J Proteome Res*, 7(9):4050–7, 2008.

L L McQuitty. Similarity analysis by reciprocal pairs for discrete and continuous data. *Educ Psychol Meas*, 26:825–31, 1966.

R Mehta, R K Jain, and S Badve. Personalized medicine: the road ahead. *Clin Breast Cancer*, 11(1):20–6, 2011.

R Meier, D R Alessi, P Cron, M Andjelković, and B A Hemmings. Mitogenic activation, phosphorylation, and nuclear translocation of protein kinase Bbeta. *J Biol Chem*, 272(48):30491–7, 1997.

C D Michener. *The Bees of the World*. Hopkins Fulfillment Service, 2nd edition, 2007.

M L Miller and N Blom. Kinase-specific prediction of protein phosphorylation sites. *Methods Mol Biol*, 527: 299–310, 2009.

M L Miller, L J Jensen, F Diella, C Jørgensen, M Tinti, L Li, M Hsiung, S A Parker, J Bordeaux, T Sicheritz-Ponten, M Olhovsky, A Pasculescu, J Alexander, S Knapp, N Blom, P Bork, S Li, G Cesareni, T Pawson, B E Turk, M B Yaffe, S Brunak, and R Linding. Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal*, 1(35):ra2, 2008.

G Minozzi, J L Williams, A Stella, F Strozzi, M Luini, M L Settles, J F Taylor, R H Whitlock, R Zanella, and H L Neibergs. Meta-analysis of two genome-wide association studies of bovine paratuberculosis. *PLoS One*, 7(3):e32578, 2012.

D Miranda-Saavedra and G J Barton. Classification and functional annotation of eukaryotic protein kinases. *Proteins*, 68(4):893–914, 2007.

D C Montgomery. *Design and Analysis of Experiments*. Wiley, 2009.

G Moreno-Hagelsieb and K Latimer. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, 24(3):319–24, 2008.

B J T Morgan and A P G Ray. Non-uniqueness and inversions in cluster analysis. *Appl Stat*, 44(1):117–34, 1995.

A M Moses, J-K Hériché, and R Durbin. Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol*, 8(2):R23, 2007.

C A Mullin, M Frazier, J L Frazier, S Ashcraft, R Simonds, D Vanengelsdorp, and J S Pettis. High levels of miticides and agrochemicals in North American apiaries: implications for honey bee health. *PLoS One*, 5 (3):e9754, 2010.

M Mulongo, T Prysliak, E Scruten, S Napper, and J Perez-Casal. In vitro infection of bovine monocytes with *Mycoplasma bovis* delays apoptosis and suppresses production of gamma interferon and tumor necrosis factor alpha but not interleukin-10. *Infect Immun*, 82(1):62–71, 2014.

D M Mutch, J Tordjman, V Pelloux, B Hanczar, C Henegar, C Poitou, N Veyrie, J-D Zucker, and K Clément. Needle and surgical biopsy techniques differentially affect adipose tissue gene expression profiles. *Am J Clin Nutr*, 89(1):51–7, 2009.

H Nakagami, N Sugiyama, K Mochida, A Daudi, Y Yoshida, T Toyoda, M Tomita, Y Ishihama, and K Shirasu. Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. *Plant Physiol*, 153(3):1161–74, 2010.

K Nakai, A Kidera, and M Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng*, 2(2):93–100, 1988.

M Navajas, A Migeon, C Alaux, M Martin-Magniette, G Robinson, J Evans, S Cros-Arteil, D Crauser, and Y Le Conte. Differential gene expression of the honey bee *Apis mellifera* associated with *Varroa destructor* infection. *BMC Genomics*, 9:301, 2008.

F Nazzi, S P Brown, D Annoscia, F Del Piccolo, G Di Prisco, P Varricchio, G Della Vedova, F Cattonaro, E Caprio, and F Pennacchio. Synergistic parasite-pathogen interactions mediated by host immunity can drive the collapse of honeybee colonies. *PLoS Pathog*, 8(6):e1002735, 2012.

G Neuberger, G Schneider, and F Eisenhaber. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol Direct*, 2:1, 2007.

J C Obenauer, L C Cantley, and M B Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–41, 2003.

L S Ojalvo, C A Whittaker, J S Condeelis, and J W Pollard. Gene expression analysis of macrophages that facilitate tumor invasion supports a role for Wnt-signaling in mediating their activity in primary mammary tumors. *J Immunol*, 184(2):702–12, 2010.

R Overbeek, M Fonstein, M D'Souza, G D Pusch, and N Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6):2896–901, 1999.

P Pal and J W Lewis. Parasite aggregations in host populations using a reformulated negative binomial model. *J Helminthol*, 78(1):57–61, 2004.

K Parikh and M P Peppelenbosch. Kinome profiling of clinical cancer specimens. *Cancer Res*, 70(7):2575–8, 2010.

R Parker, M M Guarna, A P Melathopoulos, K-M Moon, R White, E Huxter, S F Pernal, and L J Foster. Correlation of proteome-wide changes with social immunity behaviors provides insight into resistance to the parasitic mite, *Varroa destructor*, in the honey bee (*Apis mellifera*). *Genome Biol*, 13(9):R81, 2012.

H Parkinson, U Sarkans, M Shojatalab, N Abeygunawardena, S Contrino, R Coulson, A Farne, G Garcia Lara, E Holloway, M Kapushesky, P Lilja, G Mukherjee, A Oezcimen, T Rayner, P Rocca-Serra, A Sharma, S Sansone, and A Brazma. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 33(Database issue):D553–5, 2005.

H Parkinson, M Kapushesky, M Shojatalab, N Abeygunawardena, R Coulson, A Farne, E Holloway, N Kolesnykov, P Lilja, M Lukk, R Mani, T Rayner, A Sharma, E William, U Sarkans, and A Brazma. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 35(Database issue):D747–50, 2007.

H Parkinson, M Kapushesky, N Kolesnikov, G Rustici, M Shojatalab, N Abeygunawardena, H Berube, M Dylag, I Emam, A Farne, E Holloway, M Lukk, J Malone, R Mani, E Pilicheva, T F Rayner, F Rezwan, A Sharma, E Williams, X Z Bradley, T Adamusiak, M Brandizi, T Burdett, R Coulson, M Krestyaninova, P Kurnosov, E Maguire, S G Neogi, P Rocca-Serra, S-A Sansone, N Sklyar, M Zhao, U Sarkans, and A Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868–72, 2009.

H Parkinson, U Sarkans, N Kolesnikov, N Abeygunawardena, T Burdett, M Dylag, I Emam, A Farne, E Hastings, E Holloway, N Kurbatova, M Lukk, J Malone, R Mani, E Pilicheva, G Rustici, A Sharma, E Williams, T Adamusiak, M Brandizi, N Sklyar, and A Brazma. ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, 39(Database issue):D1002–4, 2011.

M P Pavlou, E P Diamandis, and I M Blasutig. The long journey of cancer biomarkers from the bench to the clinic. *Clin Chem*, 59(1):147–57, 2013.

G Pearson, F Robinson, T Beers Gibson, B E Xu, M Karandikar, K Berman, and M H Cobb. Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions. *Endocr Rev*, 22(2): 153–83, 2001.

K Pearson. III. Regression, heredity and panmixia. *Philos Trans Royal Soc London Ser A*, 187:253–318, 1986.

M P Peppelenbosch. Kinome profiling. *Scientifica*, 2012(306798), 2012.

S Pestka, S V Kotenko, G Muthukumaran, L S Izotova, J R Cook, and G Garotta. The interferon gamma (IFN-gamma) receptor: a paradigm for the multichain cytokine receptor. *Cytokine Growth Factor Rev*, 8 (3):189–206, 1997.

B Petersen, T N Petersen, P Andersen, M Nielsen, and C Lundegaard. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol*, 9:51, 2009.

D Plewczyński, A Tkacz, A Godzik, and L Rychlewski. A support vector machine approach to the identification of phosphorylation sites. *Cell Mol Biol Lett*, 10(1):73–89, 2005.

D Plewczyński, A Tkacz, L S Wyrwicz, L Rychlewski, and K Ginalski. AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. *J Mol Model*, 14(1):69–76, 2008.

E Pruesse, C Quast, K Knittel, B M Fuchs, W Ludwig, J Peplies, and F O Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, 35(21):7188–96, 2007.

C Quast, E Pruesse, P Yilmaz, J Gerken, T Schweer, P Yarza, J Peplies, and F O Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, 41(Database issue):D590–6, 2013.

J R Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1992.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.

C Raggo, M Habermehl, L A Babiuk, and P Griebel. The *in vivo* effects of recombinant bovine herpesvirus-1 expressing bovine interferon-gamma. *J Gen Virol*, 81(Pt 11):2665–73, 2000.

G Ramsay. DNA chips: state-of-the art. *Nat Biotechnol*, 16(1):40–4, 1998.

M Ressurreição, D Rollinson, A M Emery, and A J Walker. A role for p38 MAPK in the regulation of ciliary motion in a eukaryote. *BMC Cell Biol*, 12:6, 2011.

T E Rinderer, L I de Guzman, G T Delatte, J A Stelzer, V A Lancaster, V Kuznetsov, L Beaman, R Watts, and J W Harris. Resistance to the parasitic mite *Varroa destructor* in honey bees from far eastern Russia. *Apidologie*, 32:381–94, 2001.

T Ritsema and M P Peppelenbosch. Kinome profiling of sugar signaling in plants using multiple platforms. *Plant Signal Behav*, 4(12), 2009.

T Ritsema, J Joore, W van Workum, and C M J Pieterse. Kinome profiling of *Arabidopsis* using arrays of kinase consensus substrates. *Plant Methods*, 3:3, 2007.

T Ritsema, D Brodmann, S H Diks, C L Bos, V Nagaraj, C M J Pieterse, T Boller, A Wiemken, and M P Peppelenbosch. Are small GTPases signal hubs in sugar-mediated induction of fructan biosynthesis? *PLoS One*, 4(8):e6605, 2009.

T Ritsema, M van Zanten, A Leon-Reyes, L A C J Voesenek, F F Millenaar, C M J Pieterse, and A J M Peeters. Kinome profiling reveals an interaction between jasmonate, salicylate and light control of hyponastic petiole growth in *Arabidopsis thaliana*. *PLoS One*, 5(12):e14255, 2010.

D M Rocke and B Durbin. A model for measurement error for gene expression arrays. *J Comput Biol*, 8(6): 557–69, 2001.

P Rosenkranz, P Aumeier, and B Ziegelmann. Biology and control of *Varroa destructor*. *J Invertebr Pathol*, 103 Suppl 1:S96–119, 2010.

G Rustici, N Kolesnikov, M Brandizi, T Burdett, M Dylag, I Emam, A Farne, E Hastings, J Ison, M Keays, N Kurbatova, J Malone, R Mani, A Mupo, R Pedro Pereira, E Pilicheva, J Rung, A Sharma, Y A Tang, T Ternent, A Tikhonov, D Welter, E Williams, A Brazma, H Parkinson, and U Sarkans. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res*, 41(Database issue): D987–90, 2013.

G-M Ryu, P Song, K-W Kim, K-S Oh, K-J Park, and J H Kim. Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Res*, 37(4):1297–307, 2009.

I Sadowski, B-J Breitkreutz, C Stark, T-C Su, M Dahabieh, S Raithatha, W Bernhard, R Oughtred, K Dolinski, K Barreto, and M Tyers. The PhosphoGRID *Saccharomyces cerevisiae* protein phosphorylation site database: version 2.0 update. *Database (Oxford)*, 2013:bat026, 2013.

I Saha, U Maulik, S Bandyopadhyay, and D Plewczynski. Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids*, 43(2):583–94, 2012.

D Sammataro, U Gerson, and G Needham. Parasitic mites of honey bees: life history, implications, and impact. *Annu Rev Entomol*, 45:519–48, 2000.

R Sanz-Pamplona, A Berenguer, D Cordero, S Riccadonna, X Solé, M Crous-Bou, E Guinó, X Sanjuan, S Biondo, A Soriano, G Jurman, G Capella, C Furlanello, and V Moreno. Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. *PLoS One*, 7(11):e48877, 2012.

N F W Saunders and B Kobe. The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res*, 36(Web Server issue):W286–90, 2008.

N F W Saunders, R I Brinkworth, T Huber, B E Kemp, and B Kobe. Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics*, 9:245, 2008.

C F Schaefer, K Anthony, S Krupa, J Buchoff, M Day, T Hannay, and K H Buetow. PID: the Pathway Interaction Database. *Nucleic Acids Res*, 37(Database issue):D674–9, 2009.

C Scheibenbogen, U Keilholz, M Richter, R Andreesen, and W Hunstein. The interleukin-2 receptor in human monocytes and macrophages: regulation of expression and release of the alpha and beta chains (p55 and p75). *Res Immunol*, 143(1):33–7, 1992.

M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70, 1995.

Y M Schrage, I H Briaire-de Bruijn, N F C C de Miranda, J van Oosterwijk, A H M Taminiau, T van Wezel, P C W Hogendoorn, and J V M G Bovée. Kinome profiling of chondrosarcoma reveals SRC-pathway activity and dasatinib as option for treatment. *Cancer Res*, 69(15):6216–22, 2009.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. Unpublished, 2010.

U Schulze-Gahmen, H L De Bondt, and S H Kim. High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J Med Chem*, 39(23):4540–6, 1996.

D Schwartz, M F Chou, and G M Church. Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. *Mol Cell Proteomics*, 8(2):365–79, 2009.

T D Seeley. Honey bees of the Arnot Forest: a population of feral colonies persisting with *Varroa destructor* in the northeastern United States. *Apidologie*, 38:19–29, 2007.

P Senawongse, A R Dalby, and Z R Yang. Predicting the phosphorylation sites using hidden Markov models and machine learning methods. *J Chem Inf Model*, 45(4):1147–52, 2005.

J Seok, H S Warren, A G Cuenca, M N Mindrinos, H V Baker, W Xu, D R Richards, G P McDonald-Smith, H Gao, L Hennessy, C C Finnerty, C M López, S Honari, E E Moore, J P Minei, J Cuschieri, P E Bankey, J L Johnson, J Sperry, A B Nathens, T R Billiar, M A West, M G Jeschke, M B Klein, R L Gamelli, N S Gibran, B H Brownstein, C Miller-Graziano, S E Calvano, P H Mason, J P Cobb, L G Rahme, S F Lowry, R V Maier, L L Moldawer, D N Herndon, R W Davis, W Xiao, R G Tompkins, and Inflammation and Host Response to Injury, Large Scale Collaborative Research Program. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A*, 110(9):3507–12, 2013.

P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

M B Shapiro and P Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res*, 15(17):7155–74, 1987.

H-B Shen, J Yang, and K-C Chou. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, 33(1):57–67, 2007.

H Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*, 51(3):492–508, 2002.

H Shimodaira. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann Stat*, 32(6):2616–41, 2004.

S J Shin, C-F Chang, C-D Chang, S P McDonough, B Thompson, H S Yoo, and Y-F Chang. *In vitro* cellular immune responses to recombinant antigens of *Mycobacterium avium subsp. paratuberculosis*. *Infect Immun*, 73(8):5074–85, 2005.

C J A Sigrist, L Cerutti, N Hulo, A Gattiker, L Falquet, M Pagni, A Bairoch, and P Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3(3):265–74, 2002.

A H Sikkema, S H Diks, W F A den Dunnen, A ter Elst, F J G Scherpen, E W Hoving, R Ruijtenbeek, P J Boender, R de Wijn, W A Kamps, M P Peppelenbosch, and E S J M de Bont. Kinome profiling in pediatric brain tumors as a new approach for target discovery. *Cancer Res*, 69(14):5987–95, 2009.

T Sing, O Sander, N Beerenwinkel, and T Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–1, 2005.

P Singh, T L Alley, S M Wright, S Kamdar, W Schott, R Y Wilpan, K D Mills, and J H Graber. Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. *Cancer Res*, 69(24):9422–30, 2009.

S A Slaugenhaupt, A Blumenfeld, S P Gill, M Leyne, J Mull, M P Cuajungco, C B Liebert, B Chadwick, M Idelson, L Reznik, C Robbins, I Makalowska, M Brownstein, D Krappmann, C Scheidereit, C Maayan, F B Axelrod, and J F Gusella. Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. *Am J Hum Genet*, 68(3):598–605, 2001.

T F Smith and M S Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981.

G K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

G K Smyth and T Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–73, 2003.

G K Smyth, J Michaud, and H S Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–75, 2005.

B Sobolev, D Filimonov, A Lagunin, A Zakharov, O Koborova, A Kel, and V Poroikov. Functional classification of proteins based on projection of amino acid sequences: application for prediction of protein kinase substrates. *BMC Bioinformatics*, 11:313, 2010.

Z Songyang, S Blechner, N Hoagland, M F Hoekstra, H Piwnica-Worms, and L C Cantley. Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr Biol*, 4(11):973–82, 1994.

C Stark, T-C Su, A Breitkreutz, P Lourenco, M Dahabieh, B-J Breitkreutz, M Tyers, and I Sadowski. PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast *Saccharomyces cerevisiae*. *Database (Oxford)*, 2010:bap026, 2010.

J P Staveley, S A Law, A Fairbrother, and C A Menzie. A causal analysis of observed declines in managed honey bees (*Apis mellifera*). *Hum Ecol Risk Assess*, 20(2):566–91, 2014.

D Stekel. *Microarray Bioinformatics*. Cambridge University Press, 2003.

R Suzuki and H Shimodaira. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–2, 2006.

K Swaminathan, R Adamczak, A Porollo, and J Meller. Enhanced prediction of conformational flexibility and phosphorylation in proteins. *Adv Exp Med Biol*, 680:307–19, 2010.

D Swarbreck, C Wilks, P Lamesch, T Z Berardini, M Garcia-Hernandez, H Foerster, D Li, T Meyer, R Muller, L Ploetz, A Radenbaugh, S Singh, V Swing, C Tissier, P Zhang, and E Huala. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, 36(Database issue):D1009–14, 2008.

R W Sweeney, M T Collins, A P Koets, S M McGuirk, and A J Roussel. Paratuberculosis (Johne's disease) in cattle and other susceptible species. *J Vet Intern Med*, 26(6):1239–50, 2012.

P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, E Dmitrovsky, E S Lander, and T R Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–12, 1999.

D S W Tan, G V Thomas, M D Garrett, U Banerji, J S de Bono, S B Kaye, and P Workman. Biomarker-driven early clinical trials in oncology: a paradigm shift in drug development. *Cancer J*, 15(5):406–20, 2009.

Y-R Tang, Y-Z Chen, C A Canchaya, and Z Zhang. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng Des Sel*, 20(8):405–12, 2007.

A L Tarca, J E K Cooke, and J Mackay. A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data. *Bioinformatics*, 21(11):2674–83, 2005.

R Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *J Amer Stat Assoc*, 83(402):394–405, 1988.

K Tomii and M Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng*, 9(1):27–36, 1996.

B Trost and A Kusalik. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics*, 27 (21):2927–35, 2011.

B Trost, M Bickis, and A Kusalik. Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. *Immunome Res*, 3:5, 2007.

B Trost, R Arsenault, P Griebel, S Napper, and A Kusalik. DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites. *Bioinformatics*, 29(13):1693–5, 2013a.

B Trost, J Kindrachuk, P Määttänen, S Napper, and A Kusalik. PIIKA 2: An expanded, web-based platform for analysis of kinome microarray data. *PLoS One*, 8(11):e80837, 2013b.

B Trost, J Kindrachuk, E Scruten, P Griebel, A Kusalik, and S Napper. Kinotypes: stable species- and individual-specific profiles of cellular kinase activity. *BMC Genomics*, 14(1):854, 2013c.

J M Tsuruda, J W Harris, L Bourgeois, R G Danka, and G J Hunt. High-resolution linkage analyses to identify genes that influence Varroa sensitive hygiene behavior in honey bees. *PLoS One*, 7(11):e48276, 2012.

E Turner-Brannen, K-Y G Choi, R Arsenault, H El-Gabalawy, S Napper, and N Mookherjee. Inflammatory cytokines IL-32 and IL-17 have common signaling intermediates despite differential dependence on TNF-receptor 1. *J Immunol*, 186(12):7127–35, 2011.

J A Ubersax and J E Ferrell, Jr. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*, 8(7):530–41, 2007.

S Uddin, F Lekmine, A Sassano, H Rui, E N Fish, and L C Platanias. Role of Stat5 in type I interferon-signaling and transcriptional regulation. *Biochem Biophys Res Commun*, 308(2):325–30, 2003.

UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 36(Database issue): D190–5, 2008.

UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, 40(Database issue):D71–5, 2012.

UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*, 41(Database issue):D43–7, 2013.

US Department of Health and Human Services. New drug and biological drug products; evidence needed to demonstrate effectiveness of new drugs when human efficacy studies are not ethical or feasible. Technical report, Food and Drug Administration, 2002.

J W P M van Baal, S H Diks, R J A Wanders, A M Rygiel, F Milano, J Joore, J J G H M Bergman, M P Peppelenbosch, and K K Krishnadath. Comparison of kinome profiles of Barrett's esophagus with normal squamous esophagus and normal gastric cardia. *Cancer Res*, 66(24):11605–12, 2006.

L van der Maaten and G Hinton. Visualizing data using t-SNE. *J Mach Learn Res*, 9:2579–605, 2008.

P C van Weeren, K M de Bruyn, A M de Vries-Smits, J van Lint, and B M Burgering. Essential role for protein kinase B (PKB) in insulin-induced glycogen synthase kinase 3 inactivation. Characterization of dominant-negative mutant of PKB. *J Biol Chem*, 273(21):13150–6, 1998.

D Vanengelsdorp, J D Evans, C Saegerman, C Mullin, E Haubruge, B K Nguyen, M Frazier, J Frazier, D Cox-Foster, Y Chen, R Underwood, D R Tarpy, and J S Pettis. Colony collapse disorder: a descriptive study. *PLoS One*, 4(8):e6481, 2009.

R S Wallis, P Kim, S Cole, D Hanna, B B Andrade, M Maeurer, M Schito, and A Zumla. Tuberculosis biomarkers discovery: developments, needs, and challenges. *Lancet Infect Dis*, 13(4):362–72, 2013.

J Wan, S Kang, C Tang, J Yan, Y Ren, J Liu, X Gao, A Banerjee, L B M Ellis, and T Li. Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res*, 36(4):e22, 2008.

M Wang, C Li, W Chen, and C Wang. Prediction of PK-specific phosphorylation site based on information entropy. *Sci China C Life Sci*, 51(1):12–20, 2008a.

P Wang, J Sidney, C Dow, B Mothé, A Sette, and Bjoern Peters. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol*, 4(4):e1000048, 2008b.

Y-Y Wang, Si-M Chen, and H Li. Hydrogen peroxide stress stimulates phosphorylation of FoxO1 in rat aortic endothelial cells. *Acta Pharmacol Sin*, 31(2):160–4, 2010.

P Ward, L Equinet, J Packer, and C Doerig. Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics*, 5(1):79, 2004.

W R Waters, J M Miller, M V Palmer, J R Stabel, D E Jones, K A Koistinen, E M Steadham, M J Hamilton, W C Davis, and J P Bannantine. Early induction of humoral and cellular immune responses during experimental *Mycobacterium avium* subsp. *paratuberculosis* infection of calves. *Infect Immun*, 71 (9):5130–8, 2003.

J M Wettenhall and G K Smyth. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, 20(18):3705–6, 2004.

T A Whale, T K Beskorwayne, L A Babiuk, and P J Griebel. Bovine polymorphonuclear cells passively acquire membrane lipids and integral membrane proteins from apoptotic and necrotic cells. *J Leukoc Biol*, 79(6):1226–33, 2006.

R J Whittington, D J Begg, K de Silva, K M Plain, and A C Purdie. Comparative immunological and microbiological aspects of paratuberculosis as a model mycobacterial infection. *Vet Immunol Immunopathol*, 148(1-2):29–47, 2012.

B N Wilkie and B A Mallard. Genetic effects on vaccination. *Adv Vet Med*, 41:39–51, 1999.

D Witten and R Tibshirani. A comparison of fold-change and the t-statistic for microarray data analysis. Technical report, Stanford University, 2007.

I H Witten, E Frank, and M A Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition, 2011.

Y-H Wong, T-Y Lee, H-K Liang, C-M Huang, T-Y Wang, Y-H Yang, C-H Chu, H-D Huang, M-T Ko, and J-K Hwang. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res*, 35(Web Server issue):W588–94, 2007.

S-R Woo, J A Heintz, R Albrecht, R G Barletta, and C J Czuprynski. Life and death in bovine monocytes: the fate of *Mycobacterium avium* subsp. *paratuberculosis*. *Microb Pathog*, 43(2-3):106–13, 2007.

C D Wood, T M Thornton, G Sabio, R A Davis, and M Rincon. Nuclear localization of p38 MAPK in response to DNA damage. *Int J Biol Sci*, 5(5):428–37, 2009.

C-W Wu, M Livesey, S K Schmoller, E J B Manning, H Steinberg, W C Davis, M J Hamilton, and A M Talaat. Invasion and persistence of *Mycobacterium avium* subsp. *paratuberculosis* during early stages of Johne's disease in calves. *Infect Immun*, 75(5):2110–9, 2007.

Y Xue, F Zhou, M Zhu, K Ahmed, G Chen, and X Yao. GPS: a comprehensive WWW server for phosphorylation sites prediction. *Nucleic Acids Res*, 33(Web Server issue):W184–7, 2005.

Y Xue, A Li, L Wang, H Feng, and X Yao. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, 7:163, 2006.

Y Xue, J Ren, X Gao, C Jin, L Wen, and X Yao. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics*, 7(9):1598–608, 2008.

Y Xue, X Gao, J Cao, Z Liu, C Jin, L Wen, X Yao, and J Ren. A summary of computational resources for protein phosphorylation. *Curr Protein Pept Sci*, 11(6):485–96, 2010.

Y Xue, Z Liu, J Cao, Q Ma, X Gao, Q Wang, C Jin, Y Zhou, L Wen, and J Ren. GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng Des Sel*, 24(3):255–60, 2011.

M B Yaffe, G G Leparc, J Lai, T Obata, S Volinia, and L C Cantley. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol*, 19(4):348–53, 2001.

S Yamamoto, N Sakai, H Nakamura, H Fukagawa, K Fukuda, and T Takagi. INOH: ontology-based highly structured database of signal transduction pathways. *Database (Oxford)*, 2011:bar052, 2011.

F Yang, D L Stenoien, E F Strittmatter, J Wang, L Ding, M S Lipton, M E Monroe, C D Nicora, M A Gristenko, K Tang, R Fang, J N Adkins, D G Camp, 2nd, D J Chen, and R D Smith. Phosphoproteome profiling of human skin fibroblast cells in response to low- and high-dose irradiation. *J Proteome Res*, 5 (5):1252–60, 2006.

X Yang and D L Cox-Foster. Impact of an ectoparasite on the immunity and pathology of an invertebrate: evidence for host immunosuppression and viral amplification. *Proc Natl Acad Sci U S A*, 102(21):7470–5, 2005.

Y H Yang, S Dudoit, P Luu, D M Lin, V Peng, J Ngai, and T P Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2002.

Q Yao, C Bollinger, J Gao, D Xu, and J J Thelen. P(3)DB: an integrated database for plant protein phosphorylation. *Front Plant Sci*, 3:206, 2012.

Y Ye, J-H Choi, and H Tang. RAPSearch: a fast protein similarity search tool for short reads. *BMC Bioinformatics*, 12:159, 2011.

A K Yi, J G Yoon, S C Hong, T W Redford, and A M Krieg. Lipopolysaccharide and CpG DNA synergize for tumor necrosis factor-alpha production through activation of NF-kappaB. *Int Immunol*, 13(11):1391–404, 2001.

P D Yoo, Y S Ho, B B Zhou, and A Y Zomaya. SiteSeek: post-translational modification analysis using adaptive locality-effective kernel methods and new profiles. *BMC Bioinformatics*, 9:272, 2008.

Z Yu, Z Deng, H-S Wong, and L Tan. Identifying protein-kinase-specific phosphorylation sites based on the Bagging-AdaBoost ensemble approach. *IEEE Trans Nanobioscience*, 9(2):132–43, 2010.

A Zanzoni, G Ausiello, A Via, P F Gherardini, and M Helmer-Citterich. Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res*, 35(Database issue):D229–31, 2007.

A Zanzoni, D Carbajo, F Diella, P F Gherardini, A Tramontano, M Helmer-Citterich, and A Via. Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Res*, 39 (Database issue):D268–71, 2011.

O Zetterqvist, U Ragnarsson, E Humble, L Berglund, and L Engström. The minimum substrate of cyclic AMP-stimulated protein kinase, as studied by synthetic peptides representing the phosphorylatable site of pyruvate kinase (type L) of rat liver. *Biochem Biophys Res Commun*, 70(3):696–703, 1976.

D Zhang, M T Wells, C D Smart, and W E Fry. Bayesian normalization and identification for differential gene expression data. *J Comput Biol*, 12(4):391–406, 2005.

H Zhang, X Zha, Y Tan, P V Hornbeck, A J Mastrangelo, D R Alessi, R D Polakiewicz, and M J Comb. Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *J Biol Chem*, 277(42):39379–87, 2002.

J Zhang and G V Johnson. Tau protein is hyperphosphorylated in a site-specific manner in apoptotic neuronal PC12 cells. *J Neurochem*, 75(6):2346–57, 2000.

J Zhang, J Li, and H-W Deng. Identifying gene interaction enrichment for gene expression data. *PLoS One*, 4(11):e8064, 2009.

Y Zhao, H Tang, and Y Ye. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–6, 2012.

H Zheng, P Hu, D F Quinn, and Y K Wang. Phosphotyrosine proteomic study of interferon alpha signaling pathway using a combination of immunoprecipitation and immobilized metal affinity chromatography. *Mol Cell Proteomics*, 4(6):721–30, 2005.

F-F Zhou, Y Xue, G-L Chen, and X Yao. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun*, 325(4):1443–8, 2004.

T Zhou, L Sun, J Humphreys, and E J Goldsmith. Docking interactions induce exposure of activation loop in the MAP kinase ERK2. *Structure*, 14(6):1011–9, 2006.

Appendix A

Licenses to publish

In this thesis, all but one of the main chapters (3-11) contains an article published in a peer-reviewed journal. All of the journals in which these articles were published have policies that permit an author to reproduce the article in his or her thesis, and require no explicit action in order to obtain permission. For each chapter in which a published paper was included, Table A.1 gives the journal in which it was published and the URL of the journal's permissions page. This thesis also contains two figures (Figures 2.6 and 2.10) from articles of which I am not an author. Documents giving permission to reproduce these figures can be found on pages 261 and 264, respectively.

Table A.1: License information for the published articles included in this thesis.

Chapter	Journal	URL of permissions page
3–5	*Bioinformatics*	`http://www.oxfordjournals.org/access_purchase/publication_rights.html`
7	*Science Signaling*	`http://www.sciencemag.org/site/feature/contribinfo/prep/lic_info.pdf`
8	*PLOS ONE*	`http://www.plosone.org/static/license`
9	*BMC Genomics*	`http://www.biomedcentral.com/authors/license`
10	*Infection and Immunity*	`http://journals.asm.org/site/misc/ASM_Author_Statement.xhtml`
11	*Frontiers in Genetics*	`http://www.frontiersin.org/Genetics/whypublish`

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Jan 09, 2014

This is a License Agreement between Brett Trost ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3304920534202
License date	Jan 09, 2014
Licensed content publisher	Nature Publishing Group
Licensed content publication	Neuropsychopharmacology
Licensed content title	Glycogen Synthase Kinase-3: a Putative Molecular Target for Lithium Mimetic Drugs
Licensed content author	Todd D Gould, Husseini K Manji
Licensed content date	Apr 13, 2005
Volume number	30
Issue number	7
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Design and data analysis of kinome microarrays
Expected completion date	Mar 2014
Estimated size (number of pages)	230
Total	0.00 USD
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up

to a 400 words do not require NPG approval. The translation should be credited as follows:

Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

Note: For translation from the *British Journal of Cancer,* the following credit line applies.
Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Special Terms:

v1.1

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK501197754.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

Copyright Clearance Center

Confirmation Number: 11133212
Order Date: 10/24/2013

Customer Information

Customer: Brett Trost
Account Number: 3000666484
Organization: Brett Trost
Email: brett.trost@usask.ca
Phone: +1 (306) 292-5088
Payment Method: Invoice

Order Details

Bioinformatics

Billing Status: **N/A**

Order detail ID:	64092653
Article Title:	Variance stabilization applied to microarray data calibration and to the quantification of differential expression
Author(s):	Huber, W. ; et al
DOI:	10.1093/BIOINFORMATICS/18.SUPPL_
Date:	Jul 01, 2002
ISSN:	1367-4803
Publication Type:	Journal
Volume:	18
Issue:	Suppl 1
Start page:	S96
Publisher:	OXFORD UNIVERSITY PRESS

Permission Status: Granted

Permission type: Republish or display content
Type of use: reuse in a thesis/dissertation
Order License Id: 3255561462169

Requestor type	Academic/Educational institute
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Will you be translating?	No
Author of this OUP article	No
Order reference number	
Title of your thesis / dissertation	Design and data analysis of kinome microarrays
Expected completion date	Nov 2013
Estimated size(pages)	150
Publisher VAT ID	GB 125 5067 30

Note: This item was invoiced separately through our **RightsLink service.** More info **$ 0.00**

Total order items: 1 **Order Total: $0.00**

Appendix B

Supplementary material for Chapter 4

B.1 Supplementary tables

Table B.1: Performance data for PHOSFER and its variants, as well as for the comparison tools PhosPhAt and PlantPhos, using leave-one-out cross-validation. A_{ROC} values are shown, as well as sensitivity and MCC at various specificity values.

Site	Tool	A_{ROC}	Sensitivity at specificity...			MCC at specificity...		
			0.99	0.95	0.9	0.99	0.95	0.9
S	PHOSFER	0.863	0.319	0.542	0.666	0.419	0.537	0.583
	PHOSFER-NC	0.857	0.310	0.497	0.587	0.406	0.500	0.514
	PHOSFER-EW	0.857	0.301	0.506	0.599	0.398	0.507	0.521
	PHOSFER-SO	0.839	0.211	0.452	0.581	0.323	0.462	0.505
	PHOSFER-AO	0.859	0.367	0.530	0.633	0.458	0.531	0.553
	PHOSFER-AO25	0.849	0.304	0.509	0.596	0.406	0.510	0.522
	PhosPhAt	0.792	0.199	0.399	0.508	0.313	0.417	0.444
	PlantPhos	0.796	0.129	0.341	0.508	0.238	0.371	0.448
T	PHOSFER	0.773	0.157	0.296	0.491	0.268	0.332	0.414
	PHOSFER-NC	0.769	0.093	0.204	0.519	0.190	0.238	0.450
	PHOSFER-EW	0.778	0.111	0.231	0.389	0.214	0.268	0.334
	PHOSFER-SO	0.675	0.083	0.176	0.333	0.176	0.206	0.281
	PHOSFER-AO	0.789	0.185	0.352	0.491	0.297	0.383	0.414
	PHOSFER-AO25	0.760	0.111	0.259	0.454	0.214	0.296	0.393
	PhosPhAt	0.666	0.019	0.160	0.245	0.041	0.188	0.190
	PlantPhos	0.656	0.086	0.143	0.305	0.179	0.163	0.249
Y	PHOSFER	0.745	0.224	0.265	0.429	0.356	0.312	0.370
	PHOSFER-NC	0.715	0.224	0.245	0.408	0.356	0.292	0.351
	PHOSFER-EW	0.696	0.122	0.245	0.347	0.255	0.292	0.293
	PHOSFER-SO	0.635	0.082	0.082	0.204	0.206	0.085	0.142
	PHOSFER-AO	0.770	0.082	0.286	0.347	0.206	0.331	0.293
	PHOSFER-AO25	0.763	0.082	0.143	0.306	0.206	0.177	0.253
	PhosPhAt	0.609	0.000	0.184	0.245	0.000	0.224	0.185
	PlantPhos	0.655	0.042	0.104	0.188	0.146	0.120	0.118

Table B.2: Performance comparison of PHOSFER and PHOSFER-SO when using different amounts of soybean data. Leave-one-out cross-validation was used. PHOSFER75 and PHOSFER-SO75 were the same as PHOSFER and PHOSFER-SO, respectively, except that they used only 75% of the soybean training data; and similarly for the tools numbered 50 (50% of the soybean training data) and 25 (25%).

Tool	A_{ROC}	Sensitivity at specificity...			MCC at specificity...		
		0.99	0.95	0.9	0.99	0.95	0.9
PHOSFER75	0.877	0.378	0.534	0.695	0.468	0.529	0.603
PHOSFER-SO75	0.844	0.205	0.510	0.598	0.319	0.515	0.527
PHOSFER50	0.884	0.380	0.590	0.705	0.463	0.574	0.614
PHOSFER-SO50	0.840	0.241	0.470	0.548	0.344	0.481	0.476
PHOSFER25	0.903	0.434	0.518	0.747	0.507	0.521	0.659
PHOSFER-SO25	0.848	0.398	0.602	0.651	0.478	0.591	0.573

B.2 Supplementary figures

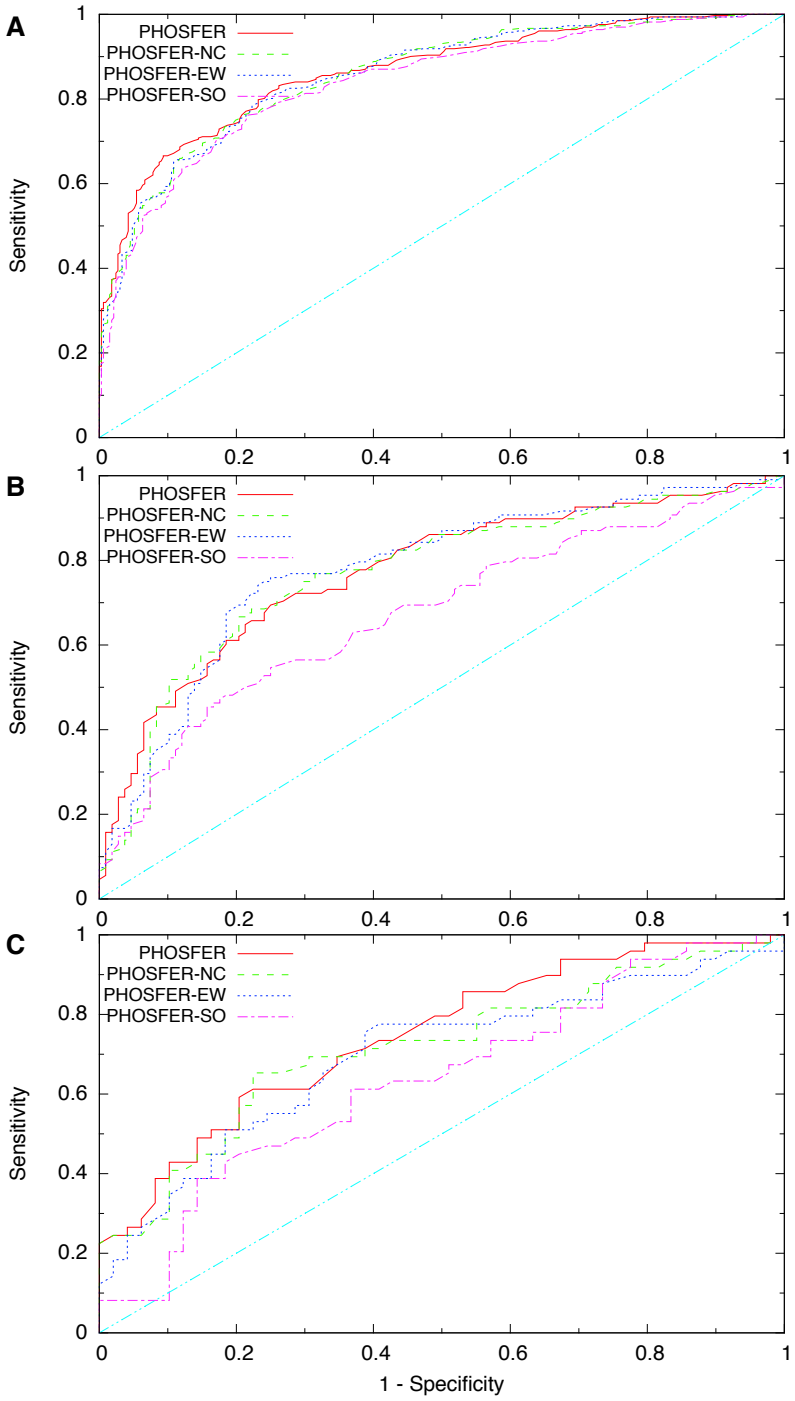


Figure B.1: ROC curves for PHOSFER and variants for (A) S phosphorylation sites, (B) T phosphorylation sites, and (C) Y phosphorylation sites. Leave-one-out cross-validation was used. The diagonal line denotes the expected performance of a tool that uses random guessing.

Appendix C

Supplementary material for Chapter 5

C.1 Detailed description of DAPPLE methodology

This document contains a detailed description of the DAPPLE methodology, complemented by a flow chart (Figure C.1) that gives a visual representation of DAPPLE's operation. To make the description easier to understand and more rigorous, symbols are used to refer to the different elements involved in the methodology, such as the target proteome, the known phosphorylation sites, and the protein corresponding to each known phosphorylation site. Many of these symbols also occur in Figure C.1, where they are rendered in blue type to better distinguish them. Also, many of the symbols correspond to column headings in the output table produced by DAPPLE. Table C.1 clarifies the relationship between these symbols and the column headings.

Let K denote the set of known phosphorylation sites. These could be derived from one or more of the following sources: PhosphoSitePlus [Hornbeck et al., 2004, 2012], Phospho.ELM [Diella et al., 2004, 2008, Dinkel et al., 2011], PhosPhAt [Heazlewood et al., 2008, Durek et al., 2010], phosphoGRID [Stark et al., 2010], P3DB [Gao et al., 2009b], or any other source of known phosphorylation sites. Let $Q \in K$ be a known phosphorylation site (i.e., sequence of amino acids) from organism Q_O, Q_L be the length of Q, Q_A be the accession number of the full protein corresponding to Q, Q_F be the sequence of the full protein with accession number Q_A, Q_C be the site (residue name and position in Q_F; e.g., Y352) of the phosphorylated residue, Q_{LTR} be the number of low-throughput references associated with Q, and Q_{HTR} be the number of high-throughput references associated with Q. Finally, let T be the target organism (the organism for which the user wants to obtain putative phosphorylation sites).

Depending on the source of a given phosphorylation site, some information may not be available. In such cases the information is recorded in the DAPPLE output table as "ND" ("not determined").

DAPPLE performs the following procedure for each $Q \in K$. In Figure C.1, the numbers shown in red correspond with the numbered steps below. Some steps of the procedure assume that $Q_L = 15$ and that the middle (eighth) residue is phosphorylated. When $Q_L < 15$, which is the case for a small portion of entries in the PhosphoSitePlus database, some of the information described below (the hit phosphorylation site (H_C), the 9-mer sequence differences (U^9), and the 9-mer non-conservative sequence differences (V^9)) cannot be determined because it is not known which residue in Q is phosphorylated. In this case, these values will be listed as "ND" in the DAPPLE output table.

1. **Obtain information from the phosphorylation database file.**

   Q_A, Q_O, Q, Q_C, Q_L, Q_{LTR}, and Q_{HTR} can be found in a single record in the database file. As mentioned above, some of this information may only be present if the data come from certain databases.

2. **Obtain the full protein sequence corresponding to the query sequence.**

   Use Q_A to retrieve Q_F in FASTA format. This record will also contain the description of this protein (Q_D).

3. **Download $\mathbf{T_P}$, the proteome of T.**

   T_P may be downloaded from any online source of protein sequence data, such as GenBank, UniProt, or IPI.

4. **Create a BLAST database comprised of the proteins in $\mathbf{T_P}$.**

   Use the `makeblastdb` program using T_P as input to create a BLAST database D_{T_P} (if D_{T_P} does not already exist).

5. **Find the most similar peptide to Q in $\mathbf{T_P}$.**

   (a) Run `blastp` using Q as the query and D_{T_P} as the database. The `-ungapped` option to `blastp` is used in order to produce an ungapped alignment.

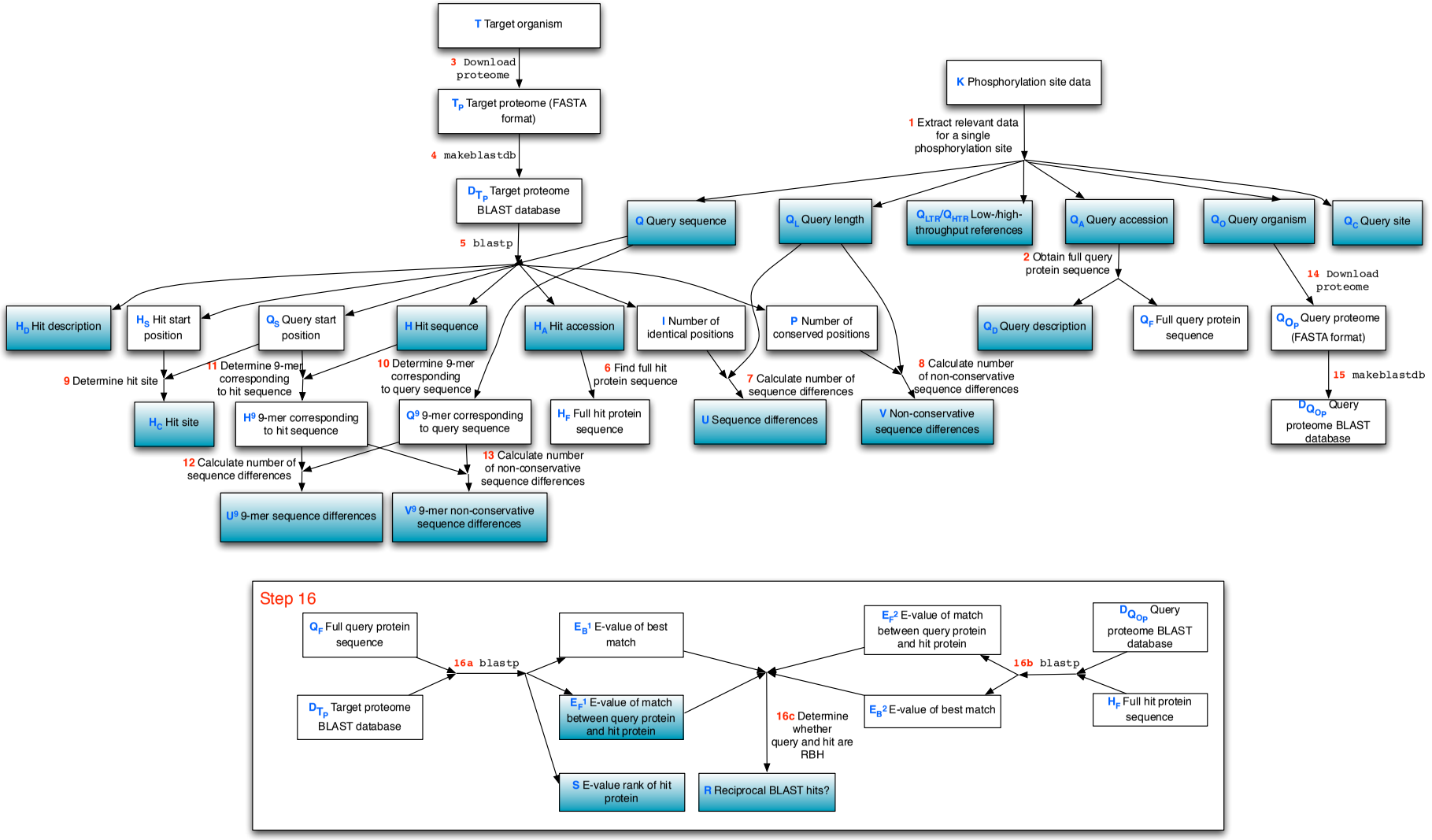


Figure C.1: Flow chart illustrating the operation of DAPPLE. The blue symbol in each box has the same meaning as the identical symbol used in this document, while red numbers correspond with the numbered steps contained herein. To achieve a cleaner-looking diagram, the rectangles labeled Q_F, D_{T_P}, $D_{Q_{O_P}}$, and H_F in the "step 16" box denote the same information as the corresponding rectangles in the main figure.

Table C.1: Correspondence between the symbols used above and the column headings in DAPPLE's output. The column headings are listed in the order that they appear in the DAPPLE output table.

Column heading	Corresponding symbol
Query accession	Q_A
Query description	Q_D
Query organism	Q_O
Query sequence	Q
Query site	Q_C
Hit site	H_C
Hit accession	H_A
Hit description	H_D
Hit sequence	H
Sequence differences	U
Non-conservative sequence differences	V
9-mer sequence differences	U^9
9-mer non-conservative sequence differences	V^9
Hit protein rank	S
Hit protein E-value	E_F^1
RBH?	R
Low-throughput references	Q_{LTR}
High-throughput references	Q_{HTR}

(b) Determine the best match H from the `blastp` search done in step 5a. Since BLAST is a local alignment program, H may be shorter than Q. The BLAST report also includes H_A (the accession number of the full protein corresponding to H), H_D (the description of that protein), I (the number of sequence identities in the alignment), P (the number of positions in the alignment that are either a match or a conservative substitution), Q_S (the query start position in the BLAST local alignment), and H_S (the hit start position). Note that Q_S is relative to Q, whereas H_S is relative to H_F (the full protein sequence having H_A as its accession number). For example, if $Q =$ ABCDEFGHIJKLMNO and the portion of Q that matches with H in the BLAST local alignment is CDEFGHIJKLMN, then $Q_S = 3$. If $H =$ CDEYGHIJKLMN and starts at position 263 in H_F, then $H_P = 263$.

6. **Obtain the full protein sequence corresponding to the hit sequence.**

   Use H_A to find H_F in T_P.

7. **Find the number of sequence differences between Q and H.**

   The number of sequence differences U is equal to $Q_L - I$.

8. **Find the number of non-conservative sequence differences between Q and H.**

   The number of non-conservative sequence differences V is equal to $Q_L - P$.

9. **Determine $\mathbf{H_C}$, the site of the phosphorylated residue in $\mathbf{H_F}$.**

   The position of this residue can be calculated using the expression $H_S - Q_S + 8$. As mentioned above, H_C cannot be determined if $Q_L < 15$.

10. **Determine the 9-amino-acid-long peptide corresponding to Q with the phosphorylated residue as its central residue.**

    The 9-amino-acid-long substring of Q with the phosphorylated residue at its center, denoted Q^9, can be found by taking the substring between indices 4 and 12, inclusive. For example, if $Q =$ ABCDEFGHIJKLMNO, then $Q^9 =$ DEFGHIJKL.

11. **Determine the 9-amino-acid-long peptide corresponding to H with the phosphorylated residue as its central residue.**

    The 9-amino-acid-long substring of H with the phosphorylated residue at its center, denoted H^9, can be found by taking the substring between indices $(5 - Q_S)$ and $(13 - Q_S)$, inclusive. For example, if $H = \text{CZEFGHIJKLMN}$ and $Q_S = 3$, then $H^9 = \text{ZEFGHIJKL}$. If H is less than nine residues long, then H^9 cannot be computed, along with U^9 and V^9 (see below).

12. **Find the number of sequence differences between $\mathbf{Q}^9$ and $\mathbf{H}^9$.**

    The number of sequence differences U^9 is the count of positions where the two residues are different in a gapless alignment between Q^9 and H^9. U^9 cannot be determined if $Q_L < 15$ or H is less than nine residues long.

13. **Find the number of non-conservative sequence differences between $\mathbf{Q}^9$ and $\mathbf{H}^9$.**

    The number of non-conservative sequence differences V^9 is the count of positions where the two residues have a non-positive score in the BLOSUM62 matrix in a gapless alignment between Q^9 and H^9. V^9 cannot be determined if $Q_L < 15$ or H is less than nine residues long.

14. **Download $\mathbf{Q_{O_P}}$, the proteome of $\mathbf{Q_O}$.**

    Q_{O_P} may be download from any online source of protein sequence data, such as GenBank, UniProt, or IPI.

15. **Create a BLAST database $\mathbf{D_{Q_{O_P}}}$ comprised of the proteins in $\mathbf{Q_{O_P}}$.**

    Use the `makeblastdb` program using Q_{O_P} as input to create a BLAST database $D_{Q_{O_P}}$ (if $D_{Q_{O_P}}$ does not already exist and Q_{O_P} exists). If no proteome exists for Q_{O_P}, then R—which denotes whether or not Q_F and H_F are reciprocal BLAST hits (see step 16)—cannot be computed.

16. **Determine whether $\mathbf{Q_F}$ and $\mathbf{H_F}$ are reciprocal BLAST hits.**

    (a) Run `blastp` using Q_F as the query and D_{T_P} as the database. Determine the E-value E_B^1 of the best BLAST hit, and also the E-value E_F^1 of the match between Q_F and H_F. Also, let S be the E-value rank of the E_F^1. In other words, if E_F^1 is the n^{th} smallest E-value, then $S = n$.

    (b) Run `blastp` using H_F as the query and $D_{Q_{O_P}}$ as the database. Determine the E-value E_B^2 of the best BLAST hit, and also the E-value E_F^2 of the match between Q_F and H_F.

    (c) Let $R =$ "yes" if Q_F and H_F are reciprocal BLAST hits, and "no" otherwise. If $E_B^1 = E_F^1$ and $E_B^2 = E_F^2$, then $R =$ "yes"; otherwise, $R =$ "no".

Example illustrating the performance of DAPPLE

The gain in efficiency using DAPPLE compared to manually performing Jalal et al. [2009]'s procedure was considerable. DAPPLE took 34 hours (elapsed time) to run on a machine with a 3.1 GHz Intel Core i5 processor and 16 GB of memory using all 179,133 unique PhosphoSitePlus-derived sites. In contrast, manually running the web-based version of BLAST and recording the results might take five minutes per peptide, or nearly 15,000 hours of labour for all of these known sites. Even the time taken to manually process a small subset of PhosphoSitePlus—say, 800 peptides, which was approximately the number used in Jalal et al. [2009]—is around 66 hours, exceeding the time required for DAPPLE to process the entire dataset.

Example illustrating the value of RBH

The usefulness of the orthologue detection procedure employed by DAPPLE can be illustrated using the following example. The human protein with accession number Q9NV56 has the annotation "MRG-binding protein". A known phosphorylation site from this protein has, as its best match in the bovine proteome, a segment of the protein with accession number E1BHM1, which has the description "Uncharacterized protein (Fragment)". These two proteins are reciprocal BLAST hits and thus orthologues—a fact that would not be possible to ascertain by comparing the annotations.

Appendix D

Supplementary material for Chapter 7

D.1 PIIKA methodology

The steps performed by PIIKA (Figure 7.1) are described in detail here. The description can be used, for example, to perform the steps of PIIKA independently of the software discussed earlier. PIIKA is implemented in the R programming language [R Development Core Team, 2006], with accessory scripts written in bash (that is, a UNIX or LINUX shell) or Perl (see Equipment). Specific R packages used in PIIKA are mentioned wherever used and can be obtained from the locations described in Equipment. Individual steps are illustrated by data samples as appropriate. In the data samples, an initial row and initial column with informative labels have been added for explanatory purposes. These may differ from the actual content of the header row and header column internally associated with the matrix by R.

D.2 Input to PIIKA

As described in the Instructions, the "create_combined_file.pl" script is used to combine the data from each individual array into the format accepted by the main PIIKA script ("piika.R"). The file produced by "create_combined_file.pl" has the format exemplified by:

peptide	protein	T1R1F	T1R1B	T1R2F	T1R2B	T2R1F	T2R1B	T2R2F	T2R2B
FAK_Y397	Q05397	33057	31091	31021	29946	43192	41861	30947	30593
FAK_Y397	Q05397	32571	31415	35434	34411	47452	46250	30716	30259
FAK_Y397	Q05397	37917	35868	44621	43545	44635	42990	31370	31069
4E-BP1_T37	Q13541	24342	30439	29591	32692	39270	42323	29800	31511
4E-BP1_T37	Q13541	25266	29416	32329	37331	37824	41222	29550	31091
4E-BP1_T37	Q13541	35696	37934	38773	43347	39216	41473	33299	34486
APE1_S289	P27695	34449	32072	29519	28403	49454	43819	32833	31121
APE1_S289	P27695	37955	35687	33782	32482	53944	48349	31895	31304
APE1_S289	P27695	35627	32936	42191	40318	45903	40279	33362	31808

where "TiRjF" stands for "Treatment i Replicate j Foreground" and "TiRjB" stands for "Treatment i Replicate j Background." In this example, the first three rows of numeric values provide the spot information for three intra-array replicates of the peptide phosphorylation location FAK_Y397 across two replicates (R1 and R2) and two treatments (T1 and T2). The remaining rows, in groups of three, contain the spot information for "4E-BP1_T37" and "APE1_S289."

D.3 Data processing before analysis

1. Background subtraction is performed on the input data. For each row and each pair of columns recording intensity values, the background intensity is subtracted from the foreground intensity. A new table is created with the results. As an example, the following matrix is the result after background subtraction is performed on the data example above:

peptide	T1R1	T1R2	T2R1	T2R2
FAK_Y397	1966	1075	1331	354
FAK_Y397	1156	1023	1202	457
FAK_Y397	2049	1076	1645	301
4E-BP1_T37	6097	−3101	−3053	-1711
4E-BP1_T37	−4150	−5002	−3398	−1541
4E-BP1_T37	−2238	−4574	−2257	−1187
APE1_S289	2377	1116	5635	1712
APE1_S289	2268	1300	5595	591
APE1_S289	2691	1873	5624	1554

In the initial row we have added for explanatory purposes "TiRj", which stands for "Treatment i Replicate j". As before, each group of three rows of numeric values provides the spot information for the three intra-array replicates (across two replicates and two treatments).

2. The resulting data is transformed with a variance stabilization (VSN) model [Huber et al., 2002]. The transformation calibrates all of the data to a positive scale while maintaining the structure within the data and alleviating variance-versus-mean dependence.

   Note: The latter problem occurs when the variances of signal intensities for individual peptides are not constant, but increase as mean intensity increases (Figures D.1, D.2, and D.3). Correction of the problem is necessary because subsequent statistical tests assume a constant variance. In addition, the data from various arrays are brought to the same scale by VSN to enable comparisons between subjects, treatments, etc. The R function "vsn2" from the vsn package is used for the transformation. It is designed for data in a table in which a single column corresponds to all of the data from a single physical microarray. This was the motivation for having intra-array replicates on separate rows in the input to this step. The wrapper function "justvsn", which is also from the vsn package, is used to simplify the use of "vsn2."

3. If there are intra-array replicates (multiple spots for individual peptides on a single array), the matrix is rearranged to have each row contain all of the replicates of a unique peptide. This is necessary because the remainder of the methodology assumes that each row of the matrix contains all replicates for a given peptide, including intra-array replicates. For example, suppose there are three intra-array replicates per peptide and that the data input to this step, after VSN transformation, are as follows:

peptide	T1R1	T1R2
FAK_Y397	11.508	11.357
FAK_Y397	11.162	11.333
FAK_Y397	11.541	11.358
4E-BP1_T37	8.157	9.113
4E-BP1_T37	8.690	8.423
4E-BP1_T37	9.426	8.557
APE1_S289	11.665	11.376
APE1_S289	11.624	11.459
APE1_S289	11.777	11.699

where the header row and column (with informative labels) have been added for explanatory purposes and "TiRj" stands for "Treatment i Replicate j". "Replicate" here could be either an inter-array or biological replicate. The dataset is then rearranged to give:

peptide	T1R1I1	T1R1I2	T1R1I3	T1R2I1	T1R2I2	T1R2I3
FAK_Y397	11.508	11.162	11.541	11.357	11.333	11.358
4E-BP1_T37	8.157	8.690	9.426	9.113	8.423	8.557
APE1_S289	11.665	11.624	11.777	11.376	11.459	11.699

where "TiRjIk" stands for "Treatment i Replicate j Intra-array Replicate k".

Note: No averaging of values is performed in steps 1 to 3. This is to maximize the number of replicates for subsequent statistical tests (χ^2-test, F-test, and t-test). Only in subsequent analysis, such as the clustering analysis in step 10, is the average for each of the peptides in a single treatment taken over the transformed replicate intensities.

Note: For sites that undergo little or no phosphorylation in a given experiment, it is not uncommon for the area surrounding a spot to undergo greater staining than the spot itself because of nonspecific interactions between the stain and the glass. This results in background intensities that are greater than the foreground intensities. Fortunately, the subsequent negative values do not present problems for the software pipeline because of the VSN transformation.

4. A χ^2-test is used to examine the variability for each peptide across technical replicate spots; that is, replicates on the same chip or multiple chips for the same subject under the same treatment [Draghici, 2003]. The results of the χ^2-tests are stored in a matrix with rows corresponding to those of the dataset. In later steps (for example, step 8), peptides with statistically significant variability may be explicitly eliminated from the dataset. For each peptide, the null hypothesis H_0 claims that there is no difference among intensities from the technical replicate spots, and the alternative hypothesis H_A states that statistically significant variation exists among them. The χ^2-test statistic (TS_1) is as follows:

$$TS_1 = \frac{(n-1)s^2}{\hat{\sigma}^2} \tag{D.1}$$

where n is the number of technical replicates for each peptide in the treatment, $s^2 = (1/n)\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance of the technical replicates for each peptide in a treatment, $\hat{\sigma}^2 = (1/M)\sum_{j=1}^{M} s_j^2$ is the mean of all the variances for the technical replicates of the M peptides in the treatment (that is, the total number of distinct peptides included in an array), and:

$$\text{P-value} = P[TS_1 > \chi^2(n-1)].$$

The peptides with P-values less than a threshold are considered to have an inconsistent pattern of phosphorylation across the technical replicates and may be eliminated in subsequent steps (steps 8, 10, or 11). When this is done, a strict confidence level (that is, 0.01) is used so that as much information as possible is retained. That is, peptides with statistically significant P-values are eliminated, so the more stringent the threshold, the fewer are discarded. The P-values are calculated with the R function "pchisq".

Note: If there are multiple technical replicate arrays, then the χ^2-test is performed for all of the replicates for a given treatment, giving a P-value for that treatment. If there are multiple biological replicate arrays, then the χ^2-test is performed separately for each array corresponding to a given treatment, and the P-value for that treatment is the minimum P-value among all these arrays.

5. One treatment may be the biological control for another treatment. Subtraction of the biological control may be useful to prepare the data for downstream analysis, such as clustering based on differences in the extent of phosphorylation. Therefore, if desired, the intensities induced by the treatments can be adjusted by subtracting the intensities of the corresponding controls. If there are multiple subjects, the biological control of the same subject is used. For example, given the following row of control and treatment information for peptide P1 in a dataset:

	BCI1	BCI2	BCI3	T1I1	T1I2	T1I3	T2I1	T2I2	T2I3
P1	4.67	3.85	4.47	3.76	4.52	3.42	4.26	4.30	4.02

this operation yields:

	T1I1′	T1I2	T1I3′	T2I1′	T2I2′	T2I3′
P1	−0.91	0.67	−1.05	−0.41	0.45	−0.45

where there is a single control (BC) for two treatments (T1 and T2), "BCIj" stands for "Biological Control Intra-array Replicate j", "TiIj" stands for "Treatment i Intra-array Replicate j", and "TiIj'" stands for "adjusted Treatment i Intra-array Replicate j". Thus BCI1 is subtracted from T1I1 and T2I1 to yield T1I1$'$ and T2I1$'$, respectively. As before, an initial row and initial column with informative labels have been added to the matrix values for explanatory purposes.

6. For each of the peptides, an F-test is used to determine whether there are statistically significant differences among the subjects under the same treatment condition [Montgomery, 2009]. This step is only applied to datasets in which there are biological replicates, and where there is a concern of variation across subjects. For example, the F-test may be important for experiments involving outbred species, including humans, where variability in responses across individuals is common. Data for peptides determined to be inconsistently phosphorylated may be eliminated in subsequent analysis (for example, in step 8). Because subtraction of the biological background may affect subject-subject variability, this step is performed after step 5.

   For a given peptide, let a be the number of subjects, n the number of intra-array replicates, N the total number of replicates for each treatment, and μ_i the mean response in the i^{th} subject for each treatment. The null hypothesis H_0 claims that $\mu_1 = \mu_2 = \ldots = \mu_a$, or the mean phosphorylation intensities elicited by the peptide among the subjects are the same, and the alternative hypothesis H_A states that not all subject means are equal. The F-statistic (TS_2) is calculated as:

$$TS_2 = \frac{MS_B}{MS_W} \tag{D.2}$$

   where

$$MS_B = \frac{SS_B}{df_B} = \frac{\sum_{i=1}^{a} n(\bar{y}_i - \bar{y})^2}{a-1} \text{ (Mean Squared Between Subjects)}$$

$$MS_W = \frac{SS_W}{df_W} = \frac{\sum_{i=1}^{a}\sum_{m=1}^{n}(y_{im} - \bar{y}_i)^2}{N-a} \text{ (Mean Squared Within Subjects).}$$

   Above, $\bar{y}_i \equiv \hat{\mu}_i$ is the sample mean for the i^{th} subject, $\bar{y} \equiv \hat{\mu}$ is the grand mean for all of the subjects, and y_{im} is the individual response of the m^{th} replicate in the i^{th} subject. Finally,

$$\text{P-value} = P[TS_2 > F(a-1, N-a)]$$

   For a given treatment, any peptide with a P-value less than a threshold for any subject is considered inconsistently phosphorylated among the subjects and may be eliminated from subsequent analysis (for example, in step 10). As with step 4, a strict confidence level (such as 0.01) is used so that as much information as possible is retained. The above calculations can be performed in R with the "aov" function.

7. For all peptides, one-sided paired t-tests are used to compare their signal intensities under two conditions, for example a treatment and a control condition [Montgomery, 2009]. This is done for all treatment-control or treatment-treatment combinations of interest. The goal is to identify those peptides for which the signal intensities are truly different under alternate conditions; that is, those peptides that are differentially phosphorylated. The paired t-test is carried out by the function "t.test" that is built into R.

   Formally, the t-test statistic (TS_3) is calculated as:

$$TS_3 = \frac{\bar{D}}{S_D/\sqrt{N}} \tag{D.3}$$

   where $\bar{D}$ is the mean of the differences between responses for a given peptide induced by two different treatments, N is the number of differences, and S_D is their standard deviation.

Finally:

$$\text{P-value (phosphorylation)} = P[TS_3 > t(N-1)]$$

and

$$\text{P-value (dephosphorylation)} = P[TS_3 < -t(N-1)].$$

Thus, each peptide has two P-values, one associated with the peptide being differentially phosphorylated and the other with the peptide being differentially dephosphorylated. The peptides with P-values less than a threshold are considered differentially (de)phosphorylated. To identify as many differentially (de)phosphorylated peptides as possible, no adjustment (as for multiple hypothesis testing) is made to the P-value, and a liberal threshold (for example, 0.1) may be used. In equation 3, N is the number of replicates per treatment. For example, if only one array was created for a single subject and there are three intra-array replicates, then $N = 3$. If there are three inter-array replicates and one subject, then $N = 9$, because there are 3 intra-array replicates per array. Finally, if there are 3 subjects and one array per subject with 3 intra-array replicates per array, then N is again 9.

Note: A paired t-test, rather than an unpaired t-test, is used here because, for a given peptide, a particular intra-array replicate for one treatment has a corresponding intra-array replicate (in the same "block" on the array) in another treatment.

Note: For some threshold T, if the P-value for differential phosphorylation of a peptide is less than T, then the P-value for dephosphorylation must be greater than T, and vice versa.

Note: The t-test is able to account for the variability among the replicates so that replicates with statistically significant P-values from the χ^2-tests will have insignificant P-values from the t-test (unless the difference in samples means is very large). However, this does not apply to datasets with multiple subjects, because significant variation for the same peptide among the subjects under the same treatment condition might be biologically meaningful, and it may confound the analysis if these peptides are treated as if they came from the same source. This was the primary motivation for the F-test in step 6.

Note: The decision to not make a P-value adjustment for multiple hypothesis testing is further discussed in Future Work.

8. The results of previous statistical tests are applied, and peptides that are differentially phosphorylated between a pair of treatments are reported. For a peptide to be deemed differentially phosphorylated, two conditions must be met. First, the P-value of the peptide from the χ^2-test must be greater than the threshold given in step 4 for both treatments. Further, if the F-test in step 6 was also applied, then the P-value of the peptide from that test must be greater than the corresponding threshold for each treatment. That is, a peptide with a χ^2-test or F-test P-value less than the corresponding threshold is not reported as differentially phosphorylated because it is deemed inconsistently phosphorylated across technical replicates or among the subjects, respectively. The second condition for a peptide to be considered differentially phosphorylated is that its P-value from either of the paired t-tests for the treatment pair is less than the threshold given in step 7.

   The strictness of the thresholds for the χ^2-test, F-test, and t-test has a direct effect on the number of peptides that are reported as differentially phosphorylated. There are fewer biomolecules represented on a kinome microarray (for example, 300) than are represented on a transcription microarray (for example, 30,000). Therefore, the thresholds are chosen so that a greater proportion of biomolecules are reported as statistically significantly different than might be the case with transcription microarrays. For the t-test, peptides with statistically significant P-values are reported, so a higher threshold yields more results. Hence, no adjustment (as for multiple hypothesis testing) is made to the P-values, and a liberal threshold (for example, 0.1) is used. For the χ^2-test and F-test, peptides with significant P-values are eliminated, so the more stringent the threshold, the fewer are discarded. Therefore, a strict confidence level (for example, 0.01) is used. In step 7, individual t-tests can be performed in parallel for various pairings of treatments. It is therefore possible that a peptide has a statistically nonsignificant P-value for one pair of treatments, but a significant P-value for another.

Note: The results of the χ^2-test and F-test are used to eliminate peptides as candidates for being differentially phosphorylated. A peptide is only eliminated if it is inconsistent for one (or both) of the treatments involved in the t-test. If it is only inconsistent for other treatments, then it is not eliminated.

Note: It is possible for a peptide to have a nonsignificant P-value for a t-test for a particular comparison between two treatments because of inconsistent intensity values, but for another combination of treatments, the intensity values are more consistent and the peptide has a significant P-value.

Note: A statistically significant χ^2-test P-value results from a large variability across replicates. This variability also results in insignificant P-values in the t-test. Hence, application of the χ^2-test results is not strictly necessary to categorize the peptide as being differentially phosphorylated, and can be bypassed for simplicity or efficiency reasons.

9. The results from the treatment-treatment variability analysis in step 7 (that is, the P-values for phosphorylation or dephosphorylation of each peptide) are reported in step 8. If there is only one treatment and a control, this often suffices for identification of differential phosphorylation. However, if there are multiple treatments relative to a single control, or multiple treatments each relative to its own control, then more complex patterns of phosphorylation may be present. For these situations, visualization of differential analysis results can facilitate the identification of patterns of differential phosphorylation across treatments.

   PIIKA makes use of a simple but effective visualization paradigm. Each peptide is represented by one small colored circle that is partitioned into two sectors (semi-circles), each of which represents a different pair of comparison treatments. For example, the left sector might be a first treatment compared with its control, whereas the right sector represents a second treatment compared to its control. A label under each circle identifies the index of the corresponding peptide in the data set. The depths of the coloration in red and green in a given sector are inversely related to the corresponding P-values for phosphorylation and dephosphorylation, respectively. For example, if the P-value for phosphorylation is 0.001, then the redness in percentage will be $100\% \times (1 - 0.001) = 99.9\%$. The same encoding is applied to dephosphorylated peptides and the extent of greenness. Thus, the combined color depths of red and green represent the phosphorylation status of each peptide in the microarray.

   The colored circles are laid out in blocks, top-to-bottom in graphical output produced by R. The first block contains peptides differentially phosphorylated in both pairs of treatments. Below that is a block of peptides differentially dephosphorylated in both pairs of treatments. Next are two sets of peptides in which one pair of treatments exhibits increased phosphorylation and the other exhibits decreased phosphorylation. Finally, peptides with inconsistent phosphorylation (as determined by the χ^2-test in step 4 or the F-test in step 6) are represented. Within the blocks in which the peptides are differentially phosphorylated in both pairs of treatments, the peptides with the most significant P-values on average for phosphorylation or dephosphorylation over the treatments being compared are presented first (going left to right and then top to bottom), followed by the less statistically significant ones. Similarly, in the blocks in which one treatment results in increased phosphorylation whereas the other yields decreased phosphorylation, peptides with the largest differences between the P-values from the treatment pairs are presented first, followed by the peptides with smaller differences. An example of the visualization for two treatment pairs is given in Figure 7.2. The visualizations are generated with the R functions "plot" (to initialize the plot), "rgb" (for coloration), and "polygon" (to draw sectors at specific coordinates to represent treatments).

   Note: The color encoding of a circle representing a peptide is specific to the treatment pairs under consideration. For example, suppose there are three treatments, a, b, and c, as well as a control, and that the spot intensities are inconsistent across subjects for treatment a, but consistent for the others. If treatments b and c versus a common control are being shown in the visualization, then the fact that treatment a is inconsistent (for this peptide) is not shown in the visualization.

   Note: It is possible to distinguish between inconsistent phosphorylation across technical replicates (the result of the χ^2-test) and inconsistent phosphorylation across subjects (the result of the F-test) in the visualization. For example, the former can be rendered in white, whereas the latter can be represented in gray. The implementation in R of such a scheme is straightforward.

Note: With more sophisticated R code, it is possible to arrange the circles in the visualization to reflect the physical layout of the array. An example is given in Figure 7.3. The peptides are grouped according to "phosphorylated for all three treatment pairs", "peptides dephosphorylated for all three treatment pairs", etc. However, now the blocks of peptides are arranged left to right and then top to bottom.

Note: It is also possible to represent more than two pairs of treatments in the visualization. In general, t treatment pairs can be represented by dividing the colored circle into t sectors. An example with $t = 3$ *is given in Figure 7.3.*

10. To further expose patterns in the kinome data, transformed peptide phosphorylation intensities are subjected to hierarchical clustering and principal component analysis (PCA). The aim is to cluster peptide response profiles across treatments or subject-treatment combinations. First, however, peptides with inconsistent intensities across technical or biological replicates are removed. Such inconsistent intensities are indicated by the P-values determined in the previous spot-spot and subject-subject variability analyses (steps 4 and 6, respectively). The same thresholds as described in step 8 are used. As opposed to the filtering in step 8, however, a peptide is removed from consideration if it is inconsistently phosphorylated for any treatment or any subject. The clustering and PCA can be across treatments or subject-treatment combinations. An average intensity is taken over the technical replicates for each treatment or subject-treatment combination. The averaged data with or without biological control subtractions is then subjected to hierarchical clustering and PCA. The dendrograms from the hierarchical clustering are augmented by heatmaps showing the averaged phosphorylation or dephosphorylation intensities.

    For hierarchical clustering, three popular combinations of linkage method and distance measurement are implemented, namely "Average Linkage + (1 - Pearson Correlation)", "Complete Linkage + Euclidean Distance", and "McQuitty + (1 - Pearson Correlation)" [McQuitty, 1966, Everitt, 1974, Hartigan, 1975, Pearson, 1986]. In general, each subject (or treatment) vector is considered as a singleton (that is, a cluster with a single element) at the initial stage of the clustering. The two most similar clusters are merged, and the distances between the newly merged clusters and the remaining clusters are updated iteratively. The calculations of similarity or distance between the clusters and the update step are algorithm-specific. The "Average Linkage + (1 - Pearson Correlation)" method is used by Eisen et al. [1998]. It takes the average over the merged (that is, the most correlated) kinome profiles and updates the distances between the merged cluster and the other clusters by recalculating the Pearson correlations between them. In "Complete Linkage + Euclidean Distance", the distance between any two clusters is considered as the Euclidean distance between the two farthest data points in the two clusters [Everitt, 1974, Hartigan, 1975]. Finally, the McQuitty method updates the distance between the two clusters in such a way that upon merging clusters C_X and C_Y into a new cluster C_{XY}, the distance between C_{XY} and each of the remaining clusters, say C_R, is calculated taking into account the sizes of C_X and C_Y [McQuitty, 1966]. These clustering methods can all be achieved with the R function "heatmap.2" from the gplots package. Input to this function includes the filtered, averaged, VSN-transformed intensity values. A particular clustering technique is specified by the arguments to the "heatmap.2" function call.

    Note: The hierarchical clustering is augmented by a heatmap, which is also generated using the R function "heatmap.2". The function converts the intensity values to statistical z-scores, and then the z-scores are encoded as color (green or red) intensities. Green usually means a value lower than the mean, whereas red represents a greater value.

    Note: PCA is a variable reduction procedure. The calculation is essentially a singular value decomposition of the centered and scaled data matrix [Mardia et al., 1979]. As a result, PCA transforms a number of possibly correlated variables into a smaller number of uncorrelated or orthogonal variables (that is, principal components). The first principal component accounts for the most variability in the data, and each succeeding component accounts for as much of the remaining variability as is possible. Usually, the first three components account for more than 50% of the variability in the data, and can be used as a set of the most important coordinates in a 3D plot to reveal the structure of the information.

    Note: The R function "prcomp" is used for PCA. A 3D plot for the PCA using the first three principal components is produced by the R function "scatterplot3d" from the package scatterplot3d. A 2D PCA

plot can be produced with the "plot" function. An example of the latter can be found in Figure D.4.

11. Although not technically part of PIIKA itself, we present here the methodology used to take output from PIIKA (the identities of the differentially phosphorylated peptides and the extent of their differential phosphorylation relative to that of peptides under control conditions) and use it to interrogate InnateDB (`http://www.innatedb.ca`) to discover known signaling pathways that are specifically influenced by the treatment under investigation [Lynn et al., 2008, Kanehisa and Goto, 2000, Kanehisa et al., 2006, 2010, Arsenault et al., 2009]. Typically, such a search requires the UniProt or GeneSymbol identifiers of the differentially phosphorylated peptides. These are readily available from the information about the kinome array, and are part of the input to PIIKA (see Materials).

    InnateDB requires fold-change (FC) values as input, with optional P-values, whereas the PIIKA methodology generates differences of transformed intensities and P-values. Therefore, to use InnateDB, the differences between the VSN-transformed intensities under the control condition and a particular treatment (or between two different treatments) are converted to ratios (that is, FC values). The formula for the VSN transformation is complex, and an inverse function is not obvious. However, an important component of the VSN transformation is calculation of a logarithm to the base 2. Hence, the conversion from the transformed intensity to the FC ratio is approximated by an exponential function (anti-logarithm).

    Peptides that show statistically significant subject-subject variability in the F-test in step 6 are removed with the threshold described in that step. In addition, peptides may be removed, if desired, based on the results of the χ^2-test; the threshold from step 4 is applied. Then, for a given treatment, the replicate transformed intensity values for each peptide are averaged. If the treatments under consideration are treatment and control, the averaging process yields $\text{average}_{\text{treatment}}$ and $\text{average}_{\text{control}}$, respectively, for each peptide. The fold-change for each peptide is then calculated as 2^d where $d = \text{average}_{\text{treatment}} - \text{average}_{\text{control}}$. This overall procedure converts the VSN-transformed values to FC ratios.

    For each of the remaining peptides in the dataset, the following is input to InnateDB: the accession number of the protein containing the peptide representing a phosphorylation site, the synthetic FC value, and a P-value from the one-sided t-test. If a peptide has a positive calculated FC value, then the P-value associated with phosphorylation is chosen. Otherwise, the P-value associated with dephosphorylation is chosen. The protein accession number was part of the information initially input to PIIKA (see Materials). If multiple peptides come from the same protein, then the protein will appear multiple times, with an individual P-value and FC value each time. InnateDB ignores column headers if given. A sample of input is given below:

protein	P-value	fold-change
Q05397	0.415457634	−1.044452633
Q13541	0.336302927	1.064849163
Q13541	0.193882405	−1.162705187
P27695	0.098999126	1.638849167

    In the sample, there are two entries for protein Q13541, the first because of the peptide with ID 4E-BP1_T37 and the second because of the peptide 4E-BP1_T46.

    Pathway analysis through InnateDB involves an interactive interface that enables specification of both P-value and FC thresholds. These thresholds specify the user's confidence in the data set and resulting pathways. InnateDB eliminates from its analysis all peptides with a P-value greater than the former threshold, or an FC value less in absolute value than the latter threshold. It is recommended that the FC threshold be set to a nonselective value, such as 1. This value is nonselective because the synthetic FC values will all be ≥ 1 or ≤ -1. This nonselectivity is a deliberate choice. Because the P-value is a calculation of how statistically significant the difference is between treatments, it is the preferred basis for determining whether a peptide should be included, rather than relying on FC. It is also recommended that the P-value threshold parameter to InnateDB be set to a liberal value such as 0.1. A more restrictive value such as 0.01 can be used, but this tends to result in very few results being reported.

InnateDB produces an extensive amount of output. The fields relevant to this analysis methodology are: (i) the pathways identified from the input proteins; (ii) the number of input proteins associated with each identified pathway; (iii) the gene symbols of the input proteins associated with each identified pathway; and (iv) a P-value for each pathway, based on the number of proteins (corresponding to input peptides) present for that pathway. Within the web interface provided by InnateDB, identified pathways can be visualized with the Cerebral plugin [Barsky et al., 2007] for the Cytoscape interaction viewer [Shannon et al., 2003]. The resultant visualizations can be downloaded to the user's computer. Examples of the resultant network visualizations are given (Figure 7.4).

Note: As in step 9, for a particular peptide it is possible for there to be statistically significant subject-subject variability (as determined by the F-test) only for treatments not under consideration. In such a case, the peptide would not be eliminated from the analysis.

Note: Peptides with large variability across their replicates will have statistically insignificant P-values in the t-test (due to a large denominator in the t-statistic), and hence will be automatically removed as a result of the threshold specified to InnateDB. Large variability across technical replicates will also result in statistically significant P-values from the χ^2-test. This is one of the reasons that filtering based on χ^2-test results is optional in this step.

D.4 Additional general notes

The organization of the input data matrix, and the restriction to disallow both inter-array replicates and biological replicates, are designed to ease the analysis in R. Alternate organizations are possible, and the restriction can be eliminated if the user is willing to devote additional R code to matrix indexing operations.

Test statistics and P-values are calculated in steps 4 and 6 for the purposes of filtering data from the analysis; however, no data are removed at those steps. Data removal is left to the subsequent steps 8 through 11. The main motivation for this is that it makes the R code for working with the matrices easier; the loops simply iterate over all 300 peptides without the need to consider exceptions. Fortunately, the presence of the inconsistently phosphorylated peptides (the peptides that would otherwise be filtered) does not harm any of the individual statistical analyses. The second reason for this design is so that each downstream step can use the results of the statistical steps in easily customizable ways. For example, in step 8, filtering based on the χ^2-test results can be optionally performed without affecting the filtering in any other step.

The visualization of step 9 automatically deals with any inconsistencies in spot intensity. Peptides with large subject-subject variability are explicitly color-coded in white. On the other hand, peptides that have large variation across replicates will have statistically insignificant P-values and hence tend to be automatically colored in brown (combined red and green). For the clustering analysis, however, these peptides need to be removed because there is no procedure that takes into account the inconsistent extent of their phosphorylation. Hence, for clustering analysis, they must be eliminated explicitly.

PIIKA is easily modified to provide information with which to search databases other than InnateDB for the discovery of known signaling pathways influenced by the treatment under investigation (step 10). For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) (`http://www.genome.jp/kegg`) [Kanehisa and Goto, 2000, Kanehisa et al., 2006, 2010] could be searched. The type and format of the information will be database-specific.

D.5 Supplementary figures

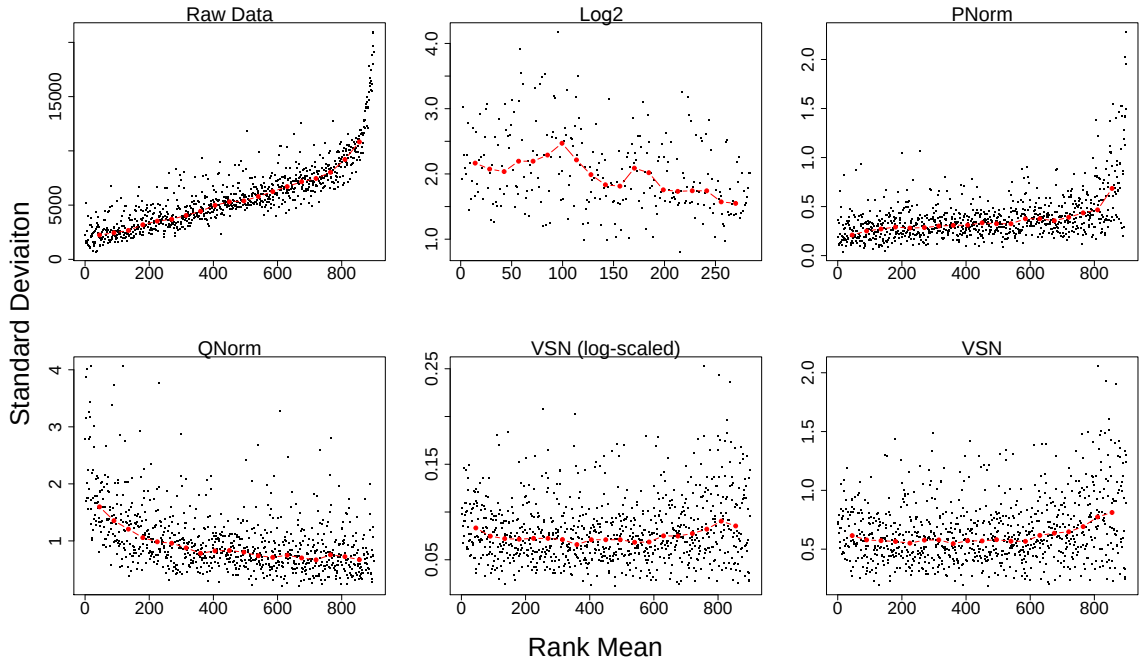


Figure D.1: Variance versus mean dependence plots before ("Raw Data") and after normalization by $\log_2$ ("Log2"), percentile normalization ("PNorm"), quantile normalization ("QNorm"), and transformation by variance stabilization ("VSN") with or without $\log_2$ scaling for the combined datasets in the case study. The rank of the mean signal intensities was plotted against the standard deviation (SD) of the corresponding peptide intensities (represented by black dots). The red dots depict the running median estimator (window-width 10%). If there is no variance-mean dependence, then the line formed by the red dots should be approximately horizontal. "Log2" refers to a simple $\log_2$ function applied after the negative values that resulted from background corrections were eliminated. The $\log_2$ function is an inbuilt function of R, and the plot was generated by the R function "meanSdPlot" from the *vsn* package [Huber et al., 2003].

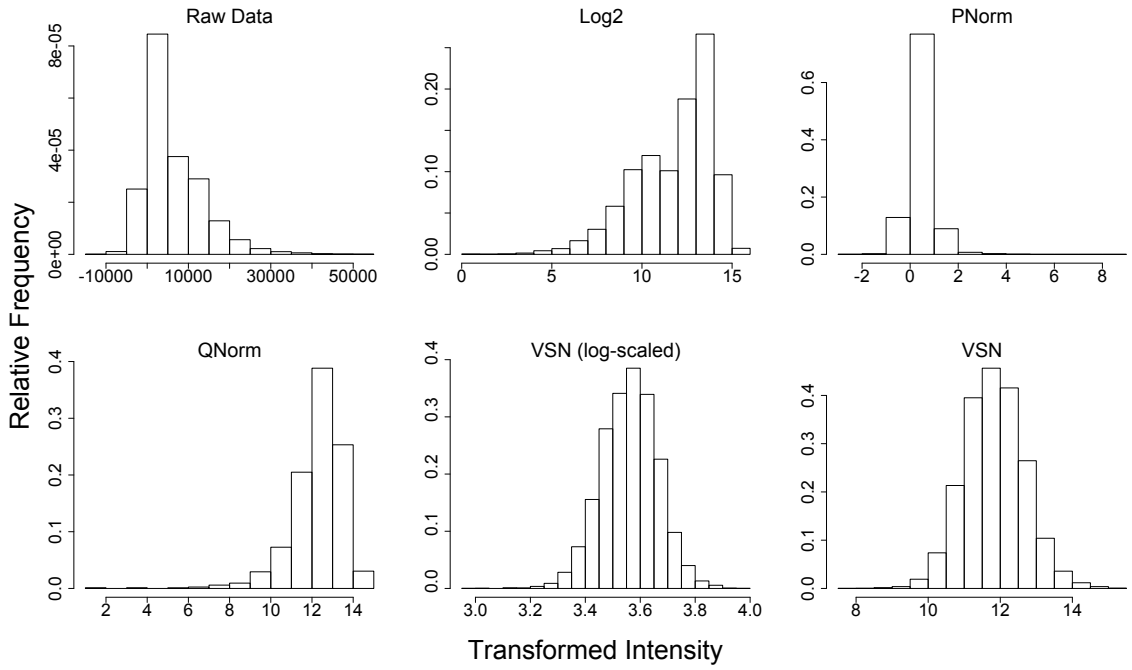


Figure D.2: Histograms of relative frequencies versus intensity before ("Raw Data") and after normalization by $\log_2$, PNorm, QNorm, or VSN with or without $\log_2$ scaling for the combined datasets in the case study. Transformations are shown as for Figure D.1. For the "Raw Data" plot, the y-axis is actual frequency.

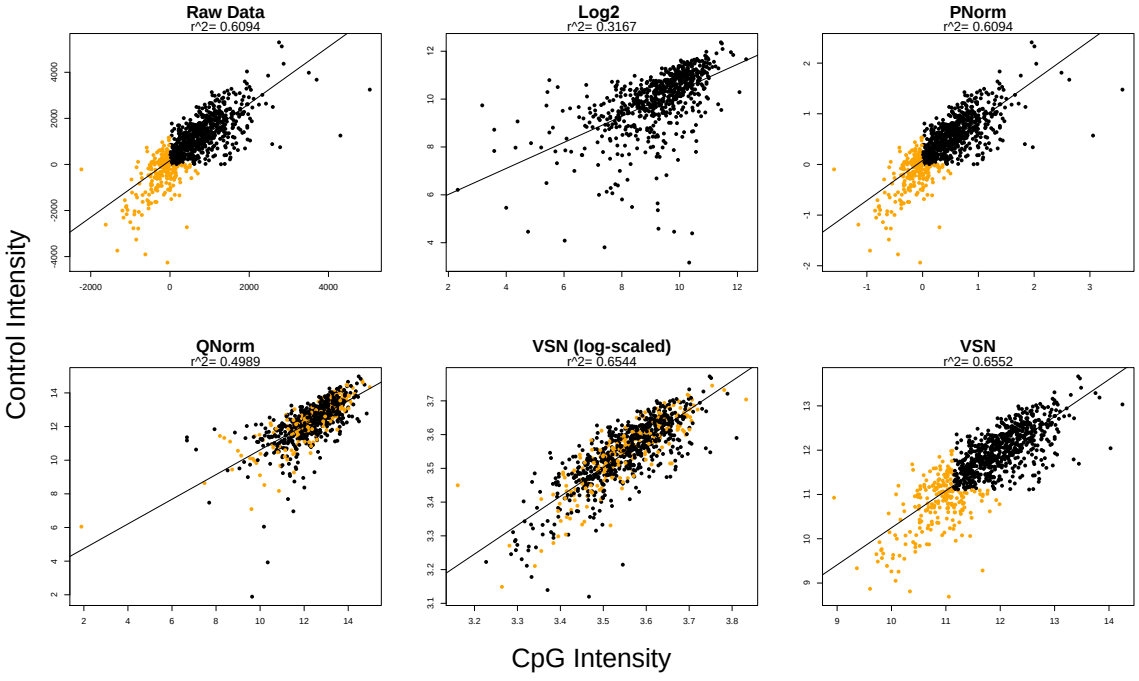


Figure D.3: Scatter plots of the signal intensities for monocytes treated with CpG oligonucleotides against the corresponding intensities from control cells treated with medium alone. The raw data were preprocessed in the following ways, as indicated: none, $\log_2$ of the positive intensities (discarding the negative ones), PNorm, QNorm, VSN (log-scaled), and VSN alone. The black and orange dots in each plot represent signal intensities after background subtraction and averaging across intra-slide replicates. If the resulting intensity for either treatment (CpG or MonoCpG) is negative, an orange dot is used. Otherwise the average intensity for both treatments is positive, and the dot is colored black. The coefficient of determination (r^2) is indicated below the title of each plot.

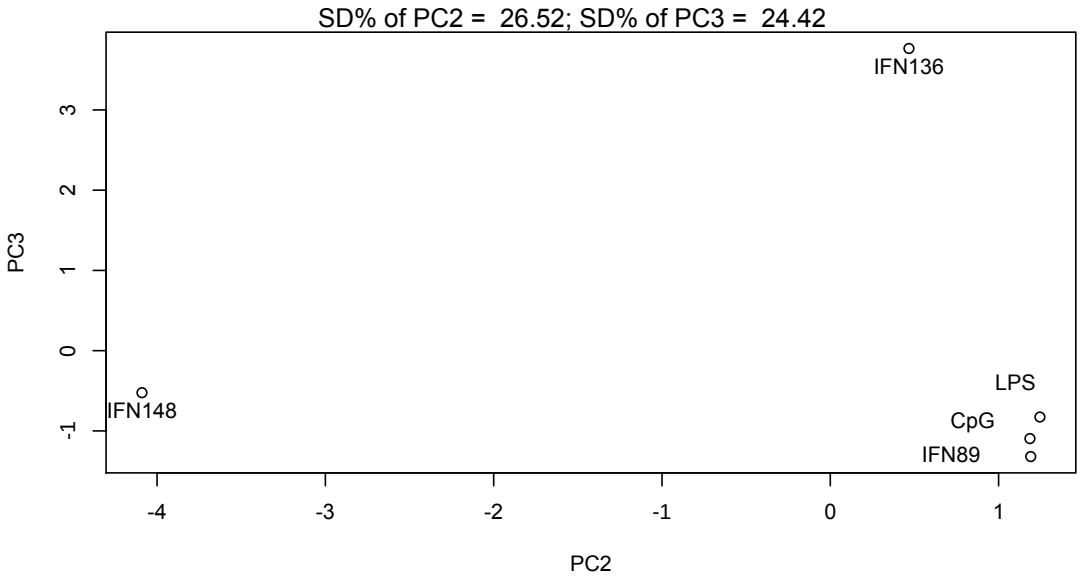


Figure D.4: Results from principal component analysis (PCA) on the intensity values from the case study. Intensity values from the three datasets were processed with our proposed PIIKA data analysis pipeline, including subtraction of biological controls, and PCA was performed on the resultant values. The second and third principal components were used for the 2D plot. The percentages of the total variability that the two PCs account for ("SD%") are displayed on the top of the plot. The data points are labeled with treatments; that is, CpG, LPS, or IFN. For the experiment involving treatment with IFN-γ, the treatment name is followed by an animal code. The R functions "prcomp" and "plot" were used for the PCA and the 2D plot, respectively.

Appendix E

Supplementary material for Chapter 8

E.1 Description of PIIKA 2 output

This document describes the output produced by PIIKA 2. The .zip file you downloaded contains several directories. Descriptions of these directories (represented by the headings below), as well as the files contained in these directories, are given below. Depending on the options you selected when you ran PIIKA, some of the directories listed below may be absent.

Several of the directories contain analyses performed after "biological subtraction", which we abbreviate in filenames as "biosub". Biological subtraction means that for each treatment-control combination specified by the user, the normalized intensity value for the control is subtracted from the normalized intensity value for the treatment for each peptide. The particular analysis being described is then performed on these subtracted values. Analyses involving biological subtraction are performed only if the user uploads a file specifying the treatment-control combinations present in the data.

Two files are contained within the top-level directory, rather than being contained within some subdirectory. These are:

- `parameters.txt`—Contains the value of the parameters used to PIIKA 2.
- `PIIKA2_output_guide.pdf`—This document.

PCA

Contains files related to Principal Component Analysis (PCA) of the various treatments.

- `PCA.txt`—A table in tab-delimited text format containing the values for the first three principal components for each treatment.
- `PCA.vrml`—Contains a 3D visualization of the PCA for each treatment in Virtual Reality Modeling Language (VRML) format. To view this file, you can use a VRML viewer such as Instant Player (`http://www.instantreality.org`).
- `PCA.vrml.legend.pdf`—Gives the colour by which each treatment is represented in `PCA.vrml`. This file will only be meaningful if your samples are named such that they indicate defined groups. For more information, see `http://saphire.usask.ca/saphire/piika/piika2_input_guide.html`.
- `PCA_PC1_PC2.pdf`—A two-dimensional scatterplot depicting the first two principal components, with the coordinates coming from `PCA.txt`.
- `PCA_PC2_PC3.pdf`—A two-dimensional scatterplot depicting the second and third principal components, with the coordinates coming from `PCA.txt`.
- `PCA_PC1_PC2_PC3.pdf`—A three-dimensional scatterplot depicting the first three principal components, with the coordinates coming from `PCA.txt`.

PCA_biosub

This directory will be present only if a file is uploaded for the "treatment-control combinations" field.

Contains files related to PCA of the various treatment-control combinations.

- `PCA_biosub.txt`—A table in tab-delimited text format containing the values for the first three principal components for each treatment-control combination.
- `PCA_biosub.vrml`—Contains a 3D visualization of the PCA for each treatment-control combination in Virtual Reality Modeling Language (VRML) format. To view this file, you can use a VRML viewer such as Instant Player (`http://www.instantreality.org`).
- `PCA_PC1_PC2.pdf`—A two-dimensional scatterplot depicting the first two principal components, with the coordinates coming from `PCA_biosub.txt`.
- `PCA_PC2_PC3.pdf`—A two-dimensional scatterplot depicting the second and third principal components, with the coordinates coming from `PCA_biosub.txt`.
- `PCA_PC1_PC2_PC3.pdf`—A three-dimensional scatterplot depicting the first three principal components, with the coordinates coming from `PCA_biosub.txt`.

biological_reproducibility

This directory will be present only if the "Perform F-test?" option is set to "Yes".

Contains files relating to the biological reproducibility of the array data (i.e., the consistency of the phosphorylation signal for each peptide in the different animals (biological replicates) for which the experiment was performed).

- `F_test_consistent_peptides.txt`—For each peptide, its value will be "TRUE" if that peptide is consistent according to the F-test for all treatments, and "FALSE" otherwise.
- `F_test_pvalues.txt`—Contains the P-value according to the F-test for each peptide for each treatment.
- `biological_reproducibility_summary.txt`—Gives the number of peptides that were biologically consistent according to the F test for each treatment, as well as the range of values and average of these values.

distances

Contains files giving numeric representations of the similarity of pairs of samples.

- `distances_euclidean.txt`—For each pair of samples, contains the Euclidean distance between that pair. Let n represent the number of peptides. Then the Euclidean distance is calculated as $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where x_i is the averaged (among all technical and biological replicates) intensity level for peptide i for the first sample, and y_i is the corresponding value for the second sample.
- `distances_pearson.txt`—For each pair of samples, contains the value (1 - Pearson correlation) for that pair. This is calculated using the `cor` function in R with `method = "pearson"`.

distances_biosub

This directory will be present only if a file is uploaded for the "treatment-control combinations" field.

Contains files giving numeric representations of the similarity of pairs of treatment-control combinations.

- `distances_biosub_euclidean.txt`—For each pair of treatment-control combinations, contains the Euclidean distance between that pair. Let n represent the number of peptides. Then the Euclidean distance is calculated as $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where x_i is the averaged (among all technical and biological replicates) intensity level for peptide i for the first treatment-control combination, and y_i is the corresponding value for the second treatment-control combination.
- `distances_biosub_pearson.txt`—For each pair of treatment-control combinations, contains the value (1 - Pearson correlation) for that pair. This is calculated using the `cor` function in R with `method = "pearson"`.

distances_significant

Contains files giving numeric representations of the similarity of pairs of samples, but taking into account only the peptides that have a statistically significant difference in phosphorylation for that pair.

- `distances_euclidean.txt`—For each pair of samples, contains the Euclidean distance between that pair, taking into account only the peptides for which the P-value from the paired t-test is less than the user-specified threshold. So that different pairs of samples can be compared, this value is then normalized by the number of significant peptides for that pair. Let n represent the number of peptides for which the paired t-test gives a P-value less than the specified threshold. Then the normalized Euclidean distance is calculated as $\frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where x_i is the averaged (among all technical and biological replicates) intensity level for peptide i for the first sample, and y_i is the corresponding value for the second sample.
- `distances_pearson.txt`—For each pair of samples, contains the value (1 - Pearson correlation) for that pair. This is calculated using the `cor` function in R with `method = "pearson"`. As with the Euclidean distance, only peptides for which the P-value from the paired t-test is less than the user-specified threshold are used in the calculation, and the resulting value is divided by the number of significant peptides so that different pairs of samples can be compared on the same scale.

distances_biosub_significant

This directory will be present only if a file is uploaded for the "treatment-control combinations" field.

Contains files giving numeric representations of the similarity of pairs of treatment-control combinations, but taking into account only the peptides that have a statistically significant difference in phosphorylation (after biological subtraction) for that pair.

- `distances_biosub_significant_euclidean.txt`—For each pair of treatment-control combinations, contains the Euclidean distance between that pair, taking into account only the peptides for which the P-value from the paired t-test is less than the user-specified threshold. So that different pairs of treatment-control combinations can be compared, this value is then normalized by the number of significant peptides for that pair. Let n represent the number of peptides for which the paired t-test gives a P-value less than the specified threshold. Then the normalized Euclidean distance is calculated as $\frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where x_i is the averaged (among all technical and biological replicates) intensity level for peptide i for the first treatment-control combination, and y_i is the corresponding value for the second treatment-control combination.
- `distances_biosub_significant_pearson.txt`—For each pair of treatment-control combinations, contains the value (1 - Pearson correlation) for that pair. This is calculated using the `cor` function in R with `method = "pearson"`. As with the Euclidean distance, only peptides for which the P-value from the paired t-test is less than the user-specified threshold are used in the calculation, and the resulting value is divided by the number of significant peptides so that different pairs of treatment-control combinations can be compared on the same scale.

hierarchical_clustering

Contains files relating to the hierarchical clustering of the samples and peptides. These files are constructed using the distance metric and linkage method chosen by the user, with the defaults being (1 - Pearson correlation) and McQuitty linkage, respectively.

- `bootstrap_dendrogram.pdf`—Contains a dendrogram depicting the hierarchical clustering of the samples, with bootstrap values as calculated using the R package `pvclust`.
- `heatmap.pdf`—Contains a heatmap wherein the columns represent samples, the rows represent peptides, and the color of the cells represent degree of up-phosphorylation (red) or down-phosphorylation (green). The top dendrogram represents the clustering of the samples, and the left dendrogram represents the clustering of the peptides.

- `heatmap.sample_dendrogram.txt`—A text-based version of the sample dendrogram depicted in the file `heatmap_euclidean_average.pdf`.

- `heatmap.peptide_dendrogram.txt`—A text-based version of the peptide dendrogram depicted in the file `heatmap_euclidean_average.pdf`.

hierarchical_clustering_biosub

This directory will be present only if a file is uploaded for the "treatment-control combinations" field.

Contains files relating to the hierarchical clustering of the treatment-control combinations and peptides. These files are constructed using the distance metric and linkage method chosen by the user, with the defaults being (1 - Pearson correlation) and McQuitty linkage, respectively.

- `bootstrap_dendrogram_biosub.pdf`—Contains a dendrogram depicting the hierarchical clustering of the treatment-control combinations, with bootstrap values as calculated using the R package `pvclust`.

- `heatmap_biosub.pdf`—Contains a heatmap wherein the columns represent treatment-control combinations, the rows represent peptides, and the color of the cells represent degree of up-phosphorylation (red) or down-phosphorylation (green) after biological subtraction. The top dendrogram represents the clustering of the treatment-control combinations, and the left dendrogram represents the clustering of the peptides.

- `heatmap_biosub.sample_dendrogram.txt`—A text-based version of the sample dendrogram depicted in the file `heatmap_biosub.pdf`.

- `heatmap_biosub.peptide_dendrogram.txt`—A text-based version of the peptide dendrogram depicted in the file `heatmap_biosub.pdf`.

intermediate_results

Contains files giving various intermediate results as the data are processed by PIIKA 2.

- `step1_raw_data.txt`—Contains the raw intensity data for each peptide for each array (foreground and background values), identical to the file uploaded by the user in the "Main input file" field.

- `step2_background_corrected.txt`—Contains the intensity value for each peptide for each array after subtracting the background from the foreground.

- `step3_vsn.txt`—Contains the normalized intensity value (normalization using the *vsn* method) for each peptide for each array.

- `step4_rearranged.txt`—Contains the same data as in `step3_vsn.txt`, except the matrix has been rearranged such that all of the intensity values corresponding to a particular peptide are in the same row.

- `step5_averages.txt`—Contains the average normalized intensity value for each treatment for each peptide.

- `step5_averages.consistent.txt`—Contains the average normalized intensity value for each treatment for each peptide that was consistent for all arrays according to the χ^2-test (if applicable), and for all animals according to the F-test (if applicable).

- `step6_biosub_averages.txt`—For each treatment-control combination, this matrix contains the subtracted value (average value for treatment minus average value for control) for each peptide. *This file will be present only if a file is uploaded for the "treatment-control combinations" field.*

- step6_biosub_averages.consistent.txt—For each treatment-control combination, this matrix contains the subtracted value (average value for treatment minus average value for control) for each peptide that was consistent for all arrays according to the χ^2-test (if applicable), and for all animals according to the F-test (if applicable). *This file will be present only if a file is uploaded for the "treatment-control combinations" field.*

scatterplots

For each pair of samples, contains a scatterplot depicting the averaged normalized intensity for each peptide for each sample in that pair.

- <sample1>_vrs_<sample2>.pdf—A scatterplot depicting the relationship between the averaged normalized intensity values for sample 1 and the averaged normalized intensity values for sample 2.

scatterplots_biosub

For each pair of treatment-control combinations, contains a scatterplot depicting the averaged normalized intensity for each peptide for each treatment-control combination in that pair.

- <treatment-control_combination1>_vrs_<treatment-control_combination2>.pdf—A scatterplot depicting the relationship between the averaged normalized intensity values for the first treatment-control combination and the averaged normalized intensity values for the second treatment-control combination.

t-tests

Contains files relating to the statistical significance of differences in phosphorylation between each treatment and control.

- <sample1>_vrs_<sample2>.all.txt—A table in tab-delimited text format giving various statistical measures of the difference in phosphorylation of each peptide in sample 1 (treatment) versus sample 2 (control). The peptides are sorted in order of increasing P-value (where this P-value is the smaller of the P-value for up-phosphorylation or down-phosphorylation). The first row contains column headings, the meanings of which are described below.
  - ID—The name of the protein from which the peptide is derived.
  - Accession—The accession number of that protein.
  - FC—The fold-change value for the peptide in the treatment versus the control.
  - P up—The P-value for up-phosphorylation in the treatment compared to the control according to the paired t-test.
  - P down—The P-value for down-phosphorylation in the treatment compared to the control according to the paired t-test.
  - Beta up—The value of β for up-phosphorylation in the treatment compared to the control.
  - Beta down—The value of β for down-phosphorylation in the treatment compared to the control.
  - Negative predictive value up—The negative predictive value for up-phosphorylation in the treatment compared to the control.
  - Negative predictive value down—The negative predictive value for down-phosphorylation in the treatment compared to the control.
- <sample1>_vrs_<sample2>.all.unsorted.txt—The same as <sample1>_vrs_<sample2>.all.txt, except not sorted by P-value.

- <sample1>_vrs_<sample2>.consistent.txt—The same as <sample1>_vrs_<sample2>.all.txt, except lists only peptides that are consistently phosphorylated in both the treatment and the control (if the χ^2-test was done), and which were consistently phosphorylated among the biological replicates for both treatment and control (if the F-test was done). *This file will be present only if one or both of the "Perform χ^2-test?" or "Perform F-test" options are set to "Yes".*

- <sample1>_vrs_<sample2>.significant.txt—The same as <sample1> _vrs_<sample2>.all.txt, except lists only peptides that have a P-value for either up-phosphorylation or down-phosphorylation less than the user-specified threshold.

- <sample1>_vrs_<sample2>.consistent_significant.txt—Contains only the peptides listed in both <sample1>_vrs_<sample2>.consistent.txt and <sample1>_vrs_<sample2>.significant.txt. *This file will be present only if one or both of the "Perform χ^2-test?" or "Perform F-test" options are set to "Yes".*

- <sample1>_vrs_<sample2>.volcano.pdf—A volcano plot, which is a scatterplot with fold-change values on the x-axis and P-values on the y-axis.

- <sample1>_vrs_<sample2>.consistent.volcano.pdf—The same as <sample1>_vrs_<sample2>.volcano.pdf, except only shows peptides listed in <sample1>_vrs_<sample2>.consistent.txt. *This file will be present only if one or both of the "Perform χ^2-test?" or "Perform F-test" options are set to "Yes".*

- <sample1>_vrs_<sample2>.not_significant.txt—Contains only the peptides not listed in <sample1> _vrs_<sample2>.significant.txt.

- <sample1>_vrs_<sample2>.positive_predictive_value.txt—Contains the positive predictive value for this treatment-control combination (which is the same for all peptides).

technical_reproducibility

This directory will be present only if the "Perform χ^2-test?" option is set to "Yes".

Contains files relating to the technical reproducibility of the array data (i.e., the consistency of the phosphorylation signal for identical peptides replicated multiple times on the same array).

- chi_square_test_consistent_peptides.txt—For each peptide, its value will be "TRUE" if that peptide is consistent according to the χ^2-test for all arrays, and "FALSE" otherwise.

- chi_square_test_pvalues.txt—Contains the P-value according to the χ^2-test for each peptide for each array.

- technical_reproducibility_summary.txt—Gives the number of peptides on each array that were technically consistent according to the χ^2-test for each array, as well as the range of values and average of these values.

random_trees

This directory will be present only if the "Perform random tree analysis?" option is set to "Yes".

Contains files related to the random tree analysis described in the main paper, which seeks to answer the question, "Do the samples cluster together better than would be expected by chance?". These files are constructed using the distance metric and linkage method chosen by the user, with the defaults being (1 - Pearson correlation) and McQuitty linkage, respectively.

- heatmap_random_<n>.averages.txt—For the nth random dendrogram, contains the randomly-rearranged matrix used to generate that dendrogram.

- `heatmap_random_<n>.pdf`—For the nth random dendrogram, contains the heatmap depicting that dendrogram.
- `heatmap_random_<n>.sample_dendrogram.txt`—For the nth random dendrogram, contains a text-based version of that dendrogram.
- `heatmap_random_tree_pvalue.txt`—Contains the P-value, which indicates the likelihood that the clustering of the actual tree (the dendrogram found in the `hierarchical_clustering` directory) was better than would be expected by chance. The P-value is calculated as the proportion of random trees that got scores equal to or greater than the score for the actual tree.
- `heatmap_random_tree_scores.txt`—Lists the score associated with each random tree.

peptide_subset_analysis

This directory will be present only if the "Perform peptide subset analysis?" option is set to "Yes".

Contains files related to the peptide subset analysis described in the main paper, which seeks to answer the question, "What subsets of the peptides give perfect or near-perfect clustering of the samples?". These files are constructed using the distance metric and linkage method chosen by the user, with the defaults being (1 - Pearson correlation) and McQuitty linkage, respectively.

- `best_set_<n>.heatmap.pdf`—Contains a heatmap generated using the n peptides found to have the best tree score.
- `best_set_<n>.peptides.txt`—Contains the n peptides found to have the best tree score.
- `best_set_<n>.sample_dendrogram.txt`—Contains a text-based version of the sample dendrogram generated using the n peptides found to have the best tree score.
- `best_set_<n>.score.txt`— contains the best tree score when using n peptides.

Appendix F

Supplementary material for Chapter 10

F.1 The dependence of false negative probabilities (values of β) on α

As described in the main paper, t-tests were used to compare the mean intensity of the treatment with the mean intensity of the control for each unique peptide on the array. We chose 0.2 as our threshold for rejecting the null hypothesis (i.e., that the means of the treatment and control are equal) for a given peptide. Peptides for which we rejected the null hypothesis were later used for pathway analysis.

We show here that, while 0.2 may seem like a liberal threshold, it is necessary to avoid high false negative probabilities (values of β) for the majority of the peptides. While we show this only for one comparison (between the proximal compartment of Animal 1 and the control compartment of Animal 1), the same trends hold for the other comparisons performed in this study.

For each peptide in the aforementioned comparison, the value of β was calculated using the R package *pwr* for four different values: 0.2 (our chosen P-value threshold), 0.1, 0.05, and 0.01. The distribution of β values for each of these values of α is given in Figure F.1.

Figure F.1 shows that using smaller values of α would result in very large values of β for a significant proportion of peptides. For instance, if $\alpha = 0.05$ is used, 227 out of the 300 peptides on the array have $\beta \geq 0.7$, whereas this is true of only 60 peptides when $\alpha = 0.2$. Therefore, using a high value of α is necessary to avoid large numbers of false negatives.

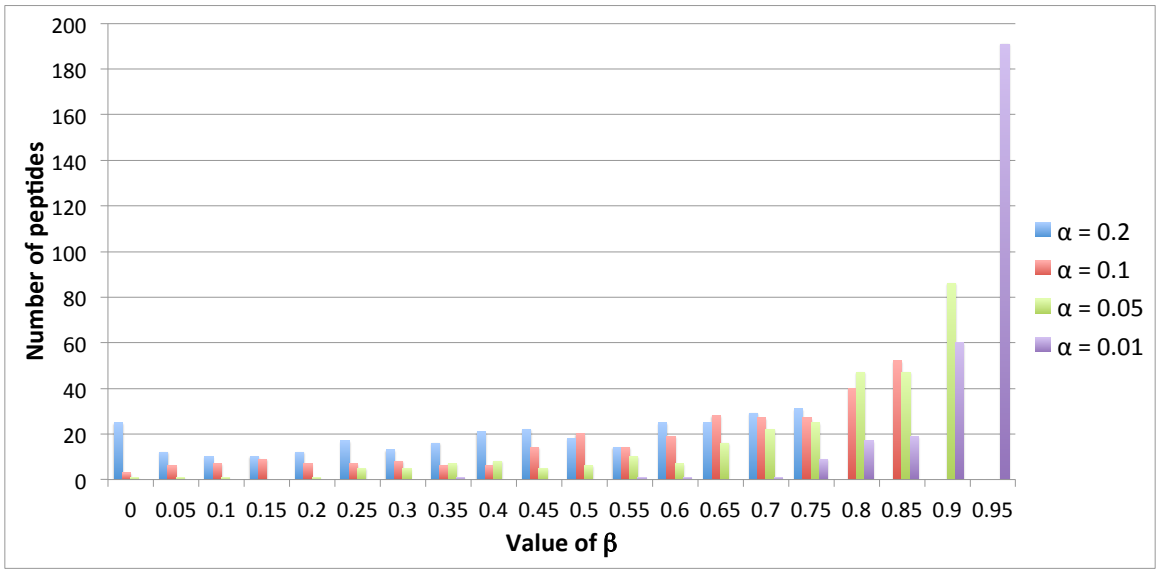


Figure F.1: Distribution of values of β for four different values of α for the comparison between the proximal compartment and the control compartment of Animal 1. The values on the x axis represent a range of values of β; for example, "0" represents values of β in the range [0,0.05). The y axis represents the number of peptides having values of β falling within that range.

Appendix G

Supplementary material for Chapter 11

G.1 Supplementary tables

Table G.1: Using sequence homology to identify honeybee phosphorylation sites. The first column indicates the number of sequence differences between a known phosphorylation site from the PhosphoSitePlus or Phospho.ELM database, and its best match in the honeybee proteome. The second column represents, for all sites in these databases, the percentage that had that number of sequence differences. The third column represents the percentage of peptides actually chosen for inclusion on the array having a given number of sequence differences.

Sequence Differences	All query peptides	Peptides on the array
0	0.6%	12.7%
1	0.8%	21.7%
2	1.1%	16.4%
3	1.2%	19.7%
4	1.4%	12.4%
5	1.6%	7.7%
6	1.6%	6.0%
7	1.3%	2.7%
8+ or no match	90.4%	0.7%

Table G.2: Pathway analysis of peptides differentially phosphorylated between resistant and susceptible uninfested bees (S88-/G4-). The columns are as follows: 1, total number of peptides; 2, number of upregulated peptides; 3, P-value for upregulation; 4, number of downregulated peptides; 5, P-value for downregulation.

Pathway Name	1	2	3	4	5
Hypoxia and p53 in the cardiovascular system	4	0	1	4	0.024
HIF-1-alpha transcription factor network	5	0	1	5	0.0089
Vegf hypoxia and angiogenesis	5	0	1	5	0.0089
Epithelial cell signaling in Helicobacter pylori infection	6	1	0.99	5	0.037
Hypoxia-inducible factor in the cardivascular system	3	0	1	3	0.062
P38 mapk signaling pathway	4	0	1	4	0.024
MAPK signaling pathway	21	8	0.98	12	0.062
Links between pyk2 and map kinases	9	2	0.99	6	0.090
Chemokine signaling pathway	12	3	0.99	8	0.047
IL2-mediated signaling events	4	0	1	4	0.024
CXCR4-mediated signaling events	6	0	1	6	0.0032
RAC1 signaling pathway	8	0	1	8	0.0004
Focal adhesion	13	2	0.99	10	0.0047
Signaling events mediated by Hepatocyte Growth Factor Receptor	9	1	0.99	8	0.0025
Endocytosis	7	7	0.013	0	1
CDC42 signaling events	11	2	0.99	9	0.0038
AndrogenReceptor	5	0	1	5	0.0089
Endothelins	5	0	1	5	0.0089
Integrin-linked kinase signaling	5	0	1	5	0.0089
ErbB2/ErbB3 signaling events	7	1	0.99	6	0.016
Class I PI3K signaling events mediated by Akt	4	0	1	4	0.024
Downstream signaling in nave CD8+ T cells	4	0	1	4	0.024
S1P2 pathway	4	0	1	4	0.024
p75(NTR)-mediated signaling	4	0	1	4	0.024
Wnt signaling pathway	6	1	0.99	5	0.037
Signaling events mediated by VEGFR1 and VEGFR2	8	2	0.98	6	0.044
VEGF signaling pathway	10	2	0.99	7	0.047
AP-1 transcription factor network	3	0	1	3	0.062
Aurora A signaling	3	0	1	3	0.062
CXCR3-mediated signaling events	3	0	1	3	0.062
Carbohydrate digestion and absorption	3	0	1	3	0.062
Cell to cell adhesion signaling	3	0	1	3	0.062
DSCAM interactions	3	0	1	3	0.062
E-cadherin signaling in the nascent adherens junction	3	0	1	3	0.062
IL6-mediated signaling events	3	0	1	3	0.062
Integrin signaling pathway	3	0	1	3	0.062
N-cadherin signaling events	3	0	1	3	0.062
Nephrin/Neph1 signaling in the kidney podocyte	3	0	1	3	0.062
Sema4D induced cell migration and growth-cone collapse	3	0	1	3	0.062
TNFalpha	19	7	0.98	11	0.068
Glucocorticoid receptor regulatory network	5	1	0.98	4	0.083
LPA receptor mediated events	5	1	0.98	4	0.083
Leukocyte transendothelial migration	5	1	0.98	4	0.083
Rac1 cell motility signaling pathway	5	1	0.98	4	0.083
Ras signaling pathway	5	1	0.98	4	0.083
Reelin signaling pathway	5	1	0.98	4	0.083
Signaling events regulated by Ret tyrosine kinase	5	1	0.98	4	0.083

Table G.2: (continued)

Wnt	11	3	0.99	7	0.088
Agrin in postsynaptic differentiation	7	1	0.99	5	0.090
Bcr signaling pathway	7	1	0.99	5	0.090
Trk receptor signaling mediated by the MAPK pathway	7	2	0.97	5	0.090
Alpha6Beta4Integrin	9	1	0.99	6	0.090
ErbB1 downstream signaling	9	3	0.96	6	0.090

Table G.3: Pathway analysis of peptides differentially phosphorylated between infested and uninfested susceptible bees (G4+/G4-). Columns are as in Table G.2.

Pathway Name	1	2	3	4	5
CDC42 signaling events	7	6	0.02	1	0.99
Signaling events mediated by Hepatocyte Growth Factor Receptor	7	6	0.02	1	0.99
Integrins in angiogenesis	4	4	0.03	0	1
Colorectal cancer	8	6	0.07	2	0.99
Pathways in cancer	18	11	0.08	7	0.97
Ctcf: first multivalent nuclear factor	3	3	0.08	0	1
Downstream signaling in nave CD8+ T cells	3	3	0.08	0	1
Endothelins	3	3	0.08	0	1
Integrin-linked kinase signaling	3	3	0.08	0	1
P38 mapk signaling pathway	3	3	0.08	0	1
Pancreatic secretion	3	3	0.08	0	1
Phagosome	3	3	0.08	0	1
Regulation of Androgen receptor activity	3	3	0.08	0	1
Regulation of retinoblastoma protein	3	3	0.08	0	1
RhoA signaling pathway	3	3	0.08	0	1
Signaling events mediated by HDAC Class III	3	3	0.08	0	1
Validated nuclear estrogen receptor alpha network	3	3	0.08	0	1
EGFR1	31	17	0.09	14	0.95
Glycolysis and Gluconeogenesis	5	4	0.11	1	0.99
RAC1 signaling pathway	5	4	0.11	1	0.99
Signaling events mediated by VEGFR1 and VEGFR2	5	4	0.11	1	0.99
Oocyte meiosis	6	0	1	6	0.03
FOXM1 transcription factor network	4	0	1	4	0.09
IL-1 signaling pathway (through p38 cascade)	4	0	1	4	0.09
IL-7	4	0	1	4	0.09
IL-9	4	0	1	4	0.09
Interleukin-1 signaling	4	0	1	4	0.09
T cell receptor signaling pathway	10	2	0.98	8	0.09
Signal transduction by L1	7	1	0.98	6	0.09

Table G.4: Pathway analysis of peptides differentially phosphorylated between infested and uninfested resistant bees (S88+/S88-). Columns are as in Table G.2.

Pathway Name	1	2	3	4	5
Pathways in cancer	10	9	0.036735	1	0.996344
MAPK signaling pathway	15	12	0.068044	3	0.98515